UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

AROLDO FERRAZ

EMPREGO DE SWIN TRANSFORMER PARA CLASSIFICAR IMAGENS RADIOGRÁFICAS DE TÓRAX E DIAGNOSTICAR COVID-19

CURITIBA
2023

AROLDO FERRAZ

EMPREGO DE SWIN TRANSFORMER PARA CLASSIFICAR IMAGENS RADIOGRÁFICAS DE TÓRAX E DIAGNOSTICAR COVID-19

Use of Swin Transformer to classify chest radiographic images and diagnose COVID-19

Dissertação apresentada como requisito para obtenção do título de Mestre em Computação Aplicada do Programa de Pós-Graduação em Computação Aplicada, pela Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Prof. Dr. Roberto Cesar Betini.

CURITIBA

2023

AROLDO FERRAZ

EMPREGO DE SWIN TRANSFORMER PARA CLASSIFICAR IMAGENS RADIOGRÁFICAS DE TÓRAX E DIAGNOSTICAR COVID-19

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Sistemas Computacionais.

Data de aprovação: 07 de agosto de 2023

Prof. Dr. Roberto Cesar Betini, Doutorado — Universidade Tecnológica Federal do Paraná
Prof. Dr. André Eugênio Lazzaretti, Doutorado — Universidade Tecnológica Federal do Paraná
Prof. Dr. Bogdan Tomoyuki Nassu, Doutorado — Universidade Tecnológica Federal do Paraná
Prof. Dr. David Menotti Gomes, Doutorado — Universidade Federal do Paraná (Ufpr)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 06/07/2023.

Dedico este trabalho a DEUS, a fonte de força e saúde que me acompanhou durante toda a jornada. À minha amada esposa, Andréia, que esteve ao meu lado, apoiando-me incondicionalmente e compreendendo os momentos de ausência. Aos meus filhos, Filippi, André e Pedro, que mesmo nos momentos em que não pude estar presente, sempre compreenderam e apoiaram meus esforços. Esta conquista é também de vocês.

AGRADECIMENTOS

Gostaria de expressar meu sincero agradecimento a todas as pessoas que fizeram parte desta importante fase da minha vida. Agradeço especialmente ao meu orientador, Prof. Dr. Roberto Cesar Betini, pela sua sabedoria e orientação durante esta trajetória. Seu apoio foi fundamental para o meu crescimento acadêmico.

Também sou grato aos meus colegas de sala, cuja colaboração e companheirismo tornaram esta experiência enriquecedora. Agradeço à Secretaria do Curso pelo suporte e cooperação ao longo do processo.

Não posso deixar de mencionar minha família, cujo apoio incondicional e incentivo foram essenciais para superar os desafios enfrentados. Sua presença e apoio em cada momento do dia a dia foram pilares fundamentais durante esta pesquisa.

Por fim, expresso minha gratidão a todos que contribuíram para o sucesso deste projeto. Sua participação e apoio foram inestimáveis para o êxito alcançado. Sou verdadeiramente grato por todo o suporte recebido ao longo deste período.

“O antifrágil se beneficia do estresse, da volatilidade e da incerteza. Cresce e se aprimora com a exposição a choques imprevisíveis, enquanto as coisas frágeis se quebram”
(Nassim Taleb, 2020).

RESUMO

FERRAZ, Aroldo. **Emprego de Swin Transformer para Classificar Imagens Radiográficas de Tórax e Diagnosticar COVID-19.** 2023. 196 f. Dissertação — Programa de Pós-Graduação em Computação Aplicada – PPGCA). Curitiba, 2023.

Segundo dados da Organização Mundial da Saúde (WHO), a pandemia de COVID-19, desde o início de 2020, já infectou mais de 770,56 milhões de pessoas em todo o mundo. Dentre os infectados, mais de 6,95 milhões de pessoas perderam a vida. A COVID-19 é uma doença extremamente contagiosa, e pode incapacitar rapidamente os sistemas de saúde se os infectados não forem diagnosticados e as medidas de isolamento e tratamento do agravamento não forem tomadas em tempo hábil. O principal teste de triagem utilizado para diagnosticar a COVID-19 foi o Real Time Reverse Transcription-Polymerase Chain Reaction-RT-PCR, que embora seja preciso e confiável, demora um tempo considerável para o diagnóstico definitivo. Desde o início da pandemia, vários pesquisadores demonstraram a viabilidade de empregar várias técnicas de Deep Learning (DL) baseadas em Redes Neurais Convolucionais (CNN), com as quais obtiveram desempenho promissor para diagnosticar a COVID-19 em radiografias e em tomografia computadorizada. No entanto, alguns autores defendem que os métodos DL existentes e baseados em CNN estão sendo desafiados por novas arquiteturas e modelos de DL. Motivados por isso, neste trabalho, propomos o uso de Swin Transformer (em vez de CNN) para triagem de COVID-19 usando as imagens de raio X de quatro datasets originais: COVID-QU-Ext Dataset, SARS-COV2-CT Dataset, HUST-19 e HCV-UFPR-COVID-19. Esses datasets foram mesclados em diferentes estratégias o que resultou em um total de 9 datasets distintos. Empregamos *transfer learning* para mitigar a questão da escassez de dados. Além disso, avaliamos se os modelos têm alto poder de generalização, em função das características aprendidas, possibilitando o treinamento em um dataset e o teste em outro, com o objetivo de avaliar se há perda significativas nos valores das métricas obtidas. Ao analisar os resultados acumulados dos experimentos em todos os datasets e estratégias, fica evidente que o Swin Transformer se destaca. Em termos de sensibilidade/recall, este apresentou um desempenho 107% e 80% superior ao ViT e CNN, respectivamente. No dataset HUST-19, o Swin Transformer alcançou a pontuação máxima (1 ou 100%) em todas as métricas e continuou apresentando desempenho notável nos outros datasets. Em comparação com o estado da arte, os resultados revelam uma performance promissora e altamente competitiva para o Swin Transformer. Além disso, foi realizado testes estatísticos que demonstraram, em vários casos, que há diferenças estatisticamente significativas nas métricas obtidas pelo Swin Transformer quando comparadas com o CNN e ViT.

Palavras-chave: Imagens de radiografia, VC, COVID-19, *deep learning, Swin Transformer*.

ABSTRACT

FERRAZ, Aroldo. **Use of Swin Transformer to Classify Chest Radiographic Images and Diagnose COVID-19**. 2023. 196 p. Dissertação — Programa de Pós-Graduação em Computação Aplicada – PPGCA). Curitiba, 2023.

According to data from the World Health Organization (WHO), since the beginning of 2020, the COVID-19 pandemic has infected over 770.56 million people worldwide. Of those infected, more than 6.95 million individuals have lost their lives. COVID-19 is an extremely contagious disease and can swiftly overwhelm healthcare systems if infections are not diagnosed, and if isolation measures and treatment for deteriorating conditions are not promptly implemented. The primary screening test used to diagnose COVID-19 has been the Real Time Reverse Transcription-Polymerase Chain Reaction (RT-PCR). Although it is accurate and reliable, it takes a significant amount of time to deliver a definitive diagnosis. Since the onset of the pandemic, several researchers have demonstrated the feasibility of employing various Deep Learning (DL) techniques based on Convolutional Neural Networks (CNN), which have shown promising performance in diagnosing COVID-19 from X-rays and CT scans. However, some authors argue that existing DL methods, based on CNN, are being challenged by newer DL architectures and models. Motivated by this, in this study, we propose the use of the Swin Transformer (instead of CNN) for COVID-19 screening using X-ray images from four original datasets: COVID-QU-Ext Dataset, SARS-COV2-CT Dataset, HUST-19, and HCV-UFPR-COVID-19. These datasets were merged using different strategies, resulting in a total of 9 distinct datasets. We employed transfer learning to address the issue of data scarcity. Moreover, we assessed whether the models have a high generalization capability, based on the learned features, allowing training on one dataset, and testing on another, aiming to evaluate if there are significant losses in the metric values obtained. Upon analyzing the cumulative results from experiments across all datasets and strategies, the Swin Transformer stands out. In terms of sensitivity/recall, it performed 107% and 80% better than ViT and CNN, respectively. In the HUST-19 dataset, the Swin Transformer achieved the maximum score (1 or 100%) for all metrics and continued to show notable performance in other datasets. Compared to the state of the art, the results reveal a promising and highly competitive performance for the Swin Transformer. Additionally, statistical tests were conducted which showed, in several instances, that there are statistically significant differences in the metrics obtained by the Swin Transformer when compared to CNN and ViT.

Keywords: X-ray images, computer vision, COVID-19, deep learning, swin transformer.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AUC	—	*Area Under the Curve*
CADe	—	*Computer-Aided Detection*
CADx	—	*Computer-Aided Diagnosis*
CCT	—	*Transformers Convolutivos Compactos*
CDE	—	*Cross-Dataset Evaluation*
CNN	—	*Convolutional Neural Networks*
COVID-19	—	*Coronavirus Disease 2019*
CXR	—	*Chest X-ray*
DISTL	—	*Distillation for Self-Supervision and Self-Training Learning*
DL	—	*Deep Learning*
DTL	—	*Deep Transfer Learning*
FN	—	Falso Negativo
FNR	—	*False Negative Rate*
FP	—	Falso Positivo
FPR	—	*False Positive Rate*
GPU	—	*Graphic Processing Units*
ILSVRC	—	*ImageNet Large Scale Visual Recognition Challenge*
ML	—	*Machine Learning*
MS	—	Mapeamento Sistemático
MSA	—	*Multi-head Self Attention*
NLP	—	*Natural Language Processing*
NPV	—	*Negative Predictive Value*
PLN	—	Processamento de Linguagem Natural
ROC	—	*Receiver Operating Characteristic Curve*
RS	—	Revisão Sistemática da Literatura
RSNA	—	*Radiological Society of North America*
RT-PCR	—	*Real Time Reverse Transcription – Polymerase Chain Reaction*
SAF	—	*Self-Attention Factorization*
SARS-CoV-2	—	*Severe Acute Respiratory Syndrome Coronavirus 2*
SOTA	—	*State of the art*
SwinT	—	*Swin Transformer Tiny*

TC	—	Tomografia Computadorizada
TFP	—	Taxa de Falsos Positivos
TN	—	*True Negative*
TNR	—	*True Negative Rate*
TP	—	*True Positive*
TPR	—	*True Positive Rate*
VC	—	Visão Computacional
ViT	—	*Vision Transformer*
VN	—	Verdadeiro Negativo
VP	—	Verdadeiro Positivo
ViT	—	*Vision Transformer*
XAI	—	*Explainable Artificial Intelligence*
WHO	—	*World Health Organization*

SUMÁRIO

1 INTRODUÇÃO

Segundo dados da Organização Mundial da Saúde — OMS (*World Health Organization*, 2023), em setembro de 2023, a COVID-19 alcançou 770.563.467 casos confirmados de infecção e 6.957.216 casos fatais em todo mundo. A COVID-19 é uma doença extremamente contagiosa, e pode paralisar rapidamente os sistemas de saúde se os infectados não forem diagnosticados e as medidas para isolar e tratar o agravamento não forem tomadas em tempo hábil. O principal teste de triagem utilizado para diagnosticar a COVID-19 foi o *Real Time Reverse Transcription – Polymerase Chain Reaction* — RT-PCR (Castro *et al*., 2020), que embora seja preciso e confiável, leva um tempo considerável para um diagnóstico definitivo.

A avaliação da existência de COVID-19 através de radiografias e tomografias de tórax é trabalhosa e demanda muito tempo dos médicos especialistas. No contexto da pandemia, o diagnóstico rápido é essencial para salvar vidas e conter o contágio. Contudo, o grande volume de casos suspeitos sobrecarrega os radiologistas, que devem examinar cada imagem. Por isso, um sistema de *Deep Learning* (DL) treinado para automatizar esse diagnóstico torna-se extremamente valioso.

Desde o início da pandemia, vários pesquisadores demonstraram a viabilidade de empregar várias técnicas de DL baseadas em *Convolutional Neural Networks* — CNN (Hassan *et al*., 2022), as quais obtiveram desempenho promissor para diagnosticar COVID-19 em radiografias (Raio X e Tomografia Computadorizada — TC). No entanto, os métodos DL baseados em CNN podem falhar em capturar o contexto global devido ao seu viés indutivo.

O viés indutivo é uma característica intrínseca dos algoritmos de aprendizado de máquina, incluindo as CNN. Ele se refere à tendência desses algoritmos em aprender padrões com base nos dados de treinamento disponíveis, o que pode levar a uma generalização incorreta para novos dados. No contexto das CNN, o viés indutivo pode resultar em modelos que não conseguem capturar o contexto global das imagens devido à sua ênfase excessiva em características específicas das imagens, o que pode afetar negativamente o desempenho em tarefas de diagnóstico médico, por exemplo.

Para auxiliar as equipes médicas propomos a implementação de um processo composto por duas etapas, a primeira com diagnóstico automatizado onde as imagens passam por um modelo de VC que realiza a primeira triagem. A submissão das imagens poderia ser realizada por meio de um portal *on-line* utilizando infraestrutura em nuvem. Durante a segunda etapa, os resultados apresentados pelo modelo na primeira etapa seriam validados por um profissional sênior especialista em radiologia. Uma rotina como essa tem o potencial de aliar a capacidade de detecção de padrões, própria dos modelos computacionais, com a de diagnóstico especializado dos médicos radiologistas e pode acelerar o trabalho e aprimorar a "acurácia" de todo o processo de diagnóstico da COVID-19.

No contexto deste trabalho propomos a utilização de um modelo de DL que vem alcançando boas métricas, quando comparado ao estado da arte na tarefa de classificação de imagens de Raio X para o diagnóstico de COVID-19, especialmente no cenário de avaliação *cross-dataset*. O modelo em questão é o *Swin Transformer*.

O *Swin Transformer* é um modelo de rede neural convolucional proposto em 2021 que usa um novo mecanismo de atenção para processar imagens de forma mais eficiente e precisa. O modelo foi desenvolvido por pesquisadores da *Microsoft Research Asia* (Liu; Shih; Zhong, 2021) e da Universidade de Tecnologia de Nanjing.

Segundo Liu *et al.* (2017b), houve avanços significativos na aplicação de modelos de VC na área médica. Essa inflexão ocorreu a partir de 2012, após as primeiras publicações sobre a utilização de CNN nesse mesmo ano (Ciresan; Meier; Schmidhuber, 2012).

Desde 2012, de acordo com Luján-García *et al.* (2020), surgiram vários estudos que deram o suporte para emprego dos modelos CNN para tarefas de imagens médicas. Entre os trabalhos podemos citar: *VGG-16* (Geng *et al.*, 2019); *Inception-v3* (Szegedy *et al.*, 2016); *Residual Networks v1* e *v2 ResNet1* e *ResNet2* (Jung; Chi, 2020; He *et al.*, 2016); *Xception Depth Separable Convolution Networks* (Chollet, 2017); *DenseNet* (Yao *et al.*, 2020), entre outros. Esses modelos baseados em CNN são frequentemente usados para implementar sistemas para tarefas de VC e para Detecção Assistida por Computador e Diagnóstico Assistido por Computador (Bakator; Radosav, 2018).

Em trabalho de revisão sistemática, Hassan *et al.* (2022) estudaram especificamente o uso de modelos CNN usados para classificação, detecção e segmentação de imagens radiográficas quanto à presença ou ausência de COVID-19. Além disso, eles mapearam conjuntos de dados que foram empregados nesses trabalhos. Quanto à tarefa de classificação, foi apontado que Yang *et al.* (2020) obtiveram uma *Area Under the Curve* (AUC) de 0,98, enquanto Castiglione *et al.* (2021) obtiveram uma precisão de 0,999.

No entanto, recentemente uma nova técnica tem se mostrado promissora para uso em tarefas de VC, o nome dela é *Vision Transformer* — ViT (Dosovitskiy *et al*., 2021). Embora o ViT seja baseado no modelo *Transformer* originalmente criado para tarefas de Processamento de Linguagem Natural (PLN, ou da sigla em inglês, NLP), conforme descrito por Wolf *et al*. (2020), o ViT tem se mostrado promissor para aplicações no domínio de imagens.

Vários autores têm demonstrado que o modelo ViT tem se posto como um modelo desafiante à hegemônica CNN em tarefas de classificação de imagens para diagnosticar a COVID-19, entre os quais se destacam: Mehboob *et al*. (2022), Park, Choi e Lee (2022), Chetoui e Akhloufi (2022), Park *et al*. (2022), Jiang *et al.* (2022), Mondal *et al.* (2022), Konwer e Prasanna (2022), Krishnan e Krishnan (2021), Zhang e Wen (2021), Rahhal *et al*. (2022), Li, Yang e Yu (2021), Dehkordi *et al.* (2021), Than *et al*. (2021).

Ainda mais recente que o ViT é o modelo *Swin Transformer* (LIU; SHIH; ZHONG, 2021), que é uma variante do primeiro. O *Swin Transformer*, como informado anteriormente, será utilizado no presente trabalho para realizar experimentos no problema de classificação de imagens de radiografias para detectar casos de COVID-19. A capacidade desse modelo de fornecer métricas de classificação melhores do que os modelos CNN e ViT será testada neste trabalho.

A principal ideia por trás do *Swin Transformer* é dividir a imagem em *patches*, depois linearizá-los e processá-los sequencialmente através de várias camadas de *Transformers*. Dessa forma, o *Swin Transformer* visa produzir representações hierárquicas desses *patches*. Essa abordagem é conhecida como processamento de imagem em *patch*[1] e permite que o modelo processe imagens de tamanho arbitrário com eficiência e escalabilidade.

[1] Do inglês, *patch-based image processing.*

O *Swin Transformer* usa um novo mecanismo de atenção chamado atenção em janelas deslizantes[2], que permite que o modelo capture informações globais e locais da imagem de forma mais precisa do que as abordagens de atenção convencionais.

Além disso, o *Swin Transformer* usa várias estratégias de otimização, como a normalização por camada[3] e a inicialização dos pesos[4], para melhorar o desempenho e a eficiência do modelo.

O *Swin Transformer* tem sido usado com sucesso em várias tarefas de VC, incluindo classificação de imagens, detecção de objetos e segmentação de imagens. O modelo tem alcançado resultados de ponta em vários *benchmarks* de referência, superando outros modelos de CNN, incluindo modelos residuais (*ResNet*) e baseados em detecção de objetos (*You Only Look Once* ou YOLO).

Em resumo, optou-se pelo *Swin Transformer* por ser uma alternativa às CNN, apresentar elevado desempenho em métricas e empregar um mecanismo de atenção inovador que processa imagens de maneira mais eficaz e escalável. Ele tem sido usado com sucesso em várias tarefas de VC e é uma das abordagens mais promissoras para processamento de imagem atualmente.

1.1 MOTIVAÇÃO

No cenário da pandemia de COVID-19, um forte requisito médico e de tratamento é a necessidade de diagnóstico rápido e preciso para desafogar os sistemas de saúde, bem como iniciar, o quanto antes, as medidas de isolamento e a medicação ou internamento necessários para os pacientes. Nesse sentido, é fundamental dispor de exames que sejam baratos e que possam ter a análise automatizada tanto quanto possível.

Além do fator tempo, que é crucial para que os pacientes sejam diagnosticados e tratados ou isolados, um sistema de diagnóstico automatizado pode aliviar a carga de trabalho dos profissionais de radiologia, ou mesmo ser utilizado em locais onde não se dispõe desses profissionais. O Brasil possui 5.570 municípios e nem todos eles dispõem de equipamentos de Raio X, porém mui-

[2] Do inglês, *sliding window attention.*
[3] Do inglês, *layer normalization.*
[4] Do inglês, *weight initialization*.

tos deles apesar de terem o equipamento, podem não ter o profissional médico radiologista disponível em regime de 24 horas por dia e sete dias por semana (24/7) para analisarem as radiografias realizadas.

Em 2019, foi publicado o estudo *O Perfil do Médico Especialista em Radiologia e Diagnóstico por Imagem no Brasil*. Esse trabalho foi realizado pelo Colégio Brasileiro de Radiologia e Diagnóstico por Imagem — CBR (SCHEFFER, 2019), em conjunto com a Faculdade de Medicina da Universidade de São Paulo (FMUSP), e trouxe dados relevantes sobre o assunto. Segundo informações coletadas, verificou-se que no Brasil há cerca de 12.868 radiologistas, o que equivale a 3% dos médicos titulados do Brasil. Além do perfil dos radiologistas no Brasil, o estudo traz informações relevantes quanto à distribuição geográfica, ajudando a compreender a realidade brasileira do setor.

A pesquisa revela que cerca de 7.608 radiologistas estavam alocados na Região Sudeste, o que representa 53,5% do total, enquanto na Região Sul, são 2.347 médicos radiologistas, representando 16,5% do total. Já no Centro-Oeste do País, são 1.262, totalizando 8,9% do contingente nacional. As regiões com menor proporção são o Nordeste e Norte. A primeira com 2.530 radiologistas (17,8%) e a segunda com 470 radiologistas, o que corresponde a 3,3%.

O estudo aponta que a densidade de radiologistas no Brasil é de 6,17 por 100 mil habitantes. No entanto, essa cifra não reflete as discrepâncias regionais, com maior presença desses profissionais nas Regiões Sudeste e Sul, em especial em metrópoles e cidades litorâneas, em contraste com a baixa densidade no interior. Para mitigar a falta desses médicos em áreas carentes, o estudo propõe o uso da telerradiologia, permitindo que laudos sejam feitos por radiologistas de localidades com maior disponibilidade desses especialistas.

Partimos da premissa de que a disponibilidade de equipamentos não seria um problema, mas sim o quantitativo de médicos radiologistas para realizar o diagnóstico adequado, preciso e em tempo integral. Nessas condições, a capacidade de trabalho dos médicos rapidamente se esgotaria, dado que teriam que trabalhar em regimes de 24/7 para darem vazão, na velocidade adequada, e elaborarem o laudo dos exames realizados em uma situação como a da pandemia de COVID-19. Ainda que os médicos radiologistas trabalhassem em regime de escala para cobrir todo o período de 24/7, não haveria profissionais suficien-

tes para dar conta de todo o volume de exames que seriam realizados no auge da pandemia de COVID-19.

A tarefa de avaliar a existência de COVID-19 apenas observando as CXR, como geralmente é feito pelos radiologistas, é trabalhoso e consome um tempo significativo, dessa forma a utilização de modelos de VC pode acelerar e automatizar essa atividade, lembrando que no cenário trazido pela pandemia de coronavírus o fator tempo gasto para o diagnóstico preciso é crucial para a preservação de vidas, bem como para conter o contágio comunitário.

A proposta do trabalho é avaliar o melhor modelo que possa ser implementado em um sistema de DL ajustado para automatizar o diagnóstico dessa doença por meio de um processo composto por duas etapas. A primeira sendo realizada pelo diagnóstico automatizado oferecido pelo *Swin Transformer*, onde as imagens passam por um modelo de VC que é encarregado de realizar a primeira triagem e previsão da existência, ou não, da doença. Em seguida, na segunda etapa, esses resultados poderiam ser validados por um profissional sênior, especialista em radiologia para emissão definitiva do laudo médico.

O processo proposto tem o potencial de aliar a capacidade de detecção de padrões, própria dos modelos de VC, com a capacidade e experiência em diagnóstico especializado dos médicos radiologistas, e pode acelerar muito o trabalho e aprimorar a acurácia de todo o processo de diagnóstico da COVID-19. Além disso, esse sistema poderia ser utilizado no interior do Brasil, tendo em vista que, como foi demonstrado pelo estudo *O Perfil do Médico Especialista em Radiologia e Diagnóstico por Imagem no Brasil* (SCHEFFER, 2019), há um déficit considerável desses profissionais no interior do país, muito embora não haja falta de equipamentos de radiografia.

A proposta mais global, então, seria aproveitar o potencial dos modelos *Swin Transformer* para auxiliar os recursos humanos (médicos radiologistas) das cidades maiores, que poderiam produzir laudos de forma mais acelerada e assim apoiar, de forma efetiva, as cidades do interior onde há escassez de profissionais radiologistas.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O presente trabalho tem como objetivo avaliar o melhor modelo para que ele que possa ser utilizado no desenvolvimento e implementação de uma solução em VC para classificar imagens de radiografias para detectar COVID-19 e pulmões saudáveis, entregando métricas computacionais superiores ao estado da arte apresentados pelos modelos CNN e ViT até o momento, e que possa ser empregado em larga escala pelos serviços de saúde.

No levantamento bibliográfico realizado foi observado que o modelo *Swin Transformer* apresenta melhores métricas de AUC, acurácia, precisão, sensibilidade, F1-score, mesmo quando são alteradas as bases de dados para a execução do teste numa uma técnica conhecida como *Cross-Dataset Evaluation* (CDE) e que será aprofundada adiante. Isso possibilita que o modelo *Swin Transformer* seja utilizado em alternativa às CNN e ao ViT.

A solução proposta pode agilizar o trabalho de laudo dos exames de Raio X executado pelos médicos tanto de capitais como em cidades do interior do território brasileiro.

O objetivo primário do trabalho é investigar a eficácia do modelo *Swin Transformer* na detecção precisa e eficiente da COVID-19 em imagens radiográficas. Especificamente, o trabalho buscou avaliar o desempenho do modelo em diferentes conjuntos de dados, comparando-o com outros modelos de DL e com o estado da arte para determinar se o modelo é capaz de superar o desempenho dos outros modelos aplicáveis ao mesmo campo de estudo.

O objetivo secundário é a análise das limitações do modelo, buscando identificar em quais cenários o modelo pode falhar e quais são suas principais vulnerabilidades, a fim de fornecer orientações para aprimorar sua implementação e uso na prática clínica.

1.2.2 Objetivos Específicos

- Customizar o algoritmo dos modelos *Swin Transformer*, na linguagem de programação *Python*, para alcançar métricas adequadas e mais elevadas que as obtidas pelas *ResNet50* (CNN) e ViT, entendidos como estado da

arte. Para tanto os modelos serão executados sob os mesmos conjuntos de dados e premissas;

- alcançar as métricas de AUC e acurácia superiores ao ViT e CNN;
- substituir o *dataset* de imagens de Raio X que foi utilizado no treinamento por outro contendo imagens de radiografias geradas com características distintas, com o objetivo de testar a capacidade de generalização do modelo *Swin Transformer*; e
- substituir o *dataset* de imagens de Raio X que foi utilizado no treinamento por outro contendo imagens de TC geradas com características distintas, com o objetivo de testar a capacidade de generalização do modelo *Swin Transformer*.

1.3 RESULTADOS OBTIDOS

Entre os resultados obtidos com o uso do *Swin Transformer* para classificar imagens de CXR e diagnosticar COVID-19 podem-se estimar os seguintes benefícios:

- melhorias no diagnóstico de COVID-19: com a utilização de um modelo de DL preciso e eficaz, é possível melhorar a taxa de acerto do diagnóstico de COVID-19, o que pode levar a um tratamento mais rápido e efetivo. Nesse sentido, Nishio *et al*. (2022) obtiveram resultados encorajadores em seus experimentos e que atestam que os modelos de VC podem superar a acurácia humana obtida nas tarefas de diagnóstico de COVID-19 em imagens de Raio X;
- redução de erros médicos: a utilização de algoritmos de DL para auxiliar no diagnóstico pode ajudar a reduzir erros médicos causados por fatores como fadiga, distração ou falta de experiência. Corroboram nesse sentido, como já destacado, o trabalho de Nishio *et al*. (2022);
- aceleração do processo de diagnóstico: com a utilização de um modelo preciso e eficiente, é possível agilizar o processo de diagnóstico, o que pode levar a um tratamento mais rápido e efetivo;
- redução de custos: com a utilização de um modelo de DL para auxiliar no diagnóstico de COVID-19, é possível reduzir custos associados a exames e procedimentos desnecessários;

- contribuição para a pesquisa científica: o projeto pode contribuir para a pesquisa científica na área de CV, DL e diagnóstico médico. Os resultados obtidos podem ser compartilhados com a comunidade científica e utilizados para aprimorar técnicas e algoritmos existentes.

1.4 ESTRUTURA DO DOCUMENTO

Este trabalho segue a seguinte estrutura: no Capítulo 2 — Revisão da Literatura —, é realizada uma análise abrangente do tema; no Capítulo 3 — Materiais e Métodos —, são descritos os recursos e as abordagens utilizadas na pesquisa; no Capítulo 4 — Experimentos, Resultados e Discussão, os resultados obtidos são apresentados e discutidos, comparando-os com os estudos revisados no Capítulo 2, que representam o estado da arte. Por fim, no Capítulo 5 — Conclusão e Trabalhos Futuros —, são apresentadas as considerações finais da dissertação, incluindo a descrição dos objetivos alcançados, os resultados obtidos, as limitações e ameaças relacionadas aos resultados da pesquisa. Além disso, são destacados os problemas em aberto e sugeridas possíveis direções para pesquisas futuras.

2 REVISÃO DA LITERATURA

Segundo recomendações de importantes autores, é de fundamental relevância para a produção cientifica a realização de Mapeamento Sistemático, bem como da Revisão Sistemática da Literatura. Dentre os autores podemos citar Nakagawa *et al.* (2017), que trazem um fluxo de trabalho prático sobre como elaborar tanto a Revisão Sistemática da Literatura, quanto o Mapeamento Sistemático. Essa obra foi utilizada como referência no presente trabalho.

2.1 FUNDAMENTAÇÃO TEÓRICA

Antes de aprofundarmos na revisão da literatura e com o objetivo de melhorar o entendimento da área, foram propostas as seções abaixo que trazem alguns conceitos e terminologias que são largamente utilizadas na área de VC, portanto precisam ficar claramente fixadas para que o entendimento do restante do trabalho não seja prejudicado.

2.1.1 Métricas que Serão Utilizadas no Presente Trabalho

No contexto da implementação, avaliação e aprimoramento de modelos de *Machine Learning* (ML) e DL é necessário entender, com maior profundidade, alguns conceitos muito importantes que são aplicáveis à avaliação dos modelos de classificação em geral. Esses conceitos são comumente chamados de métricas.

De acordo com Ferrari e Silva (2017), em problemas de classificação binária, predições podem ter quatro possíveis classes:

1. Verdadeiro Positivo (VP) quando o método diz que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva;
2. Verdadeiro Negativo (VN) quando o método diz que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa;
3. Falso Positivo (FP) quando o método diz que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa;
4. Falso Negativo (FN) quando o método diz que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva.

Supondo uma classificação para as classes COVID-19 (1 ou *true*) e Normal (0 ou *false*), temos o seguinte:

- Verdadeiro Positivo (VP): é quando uma imagem de radiografias é prevista como sendo COVID-19 e ela de fato pertence à classe COVID-19.
- Verdadeiro Negativo (VN): é quando uma imagem de radiografias é prevista como sendo Normal e ela de fato pertence à classe Normal.
- Falso Positivo (FP): é quando uma imagem de radiografias é prevista como sendo COVID-19, mas na realidade ela pertence à classe Normal.
- Falso Negativo (FN): é quando uma imagem de radiografias é prevista como sendo Normal, mas na realidade ela pertence à classe COVID-19.

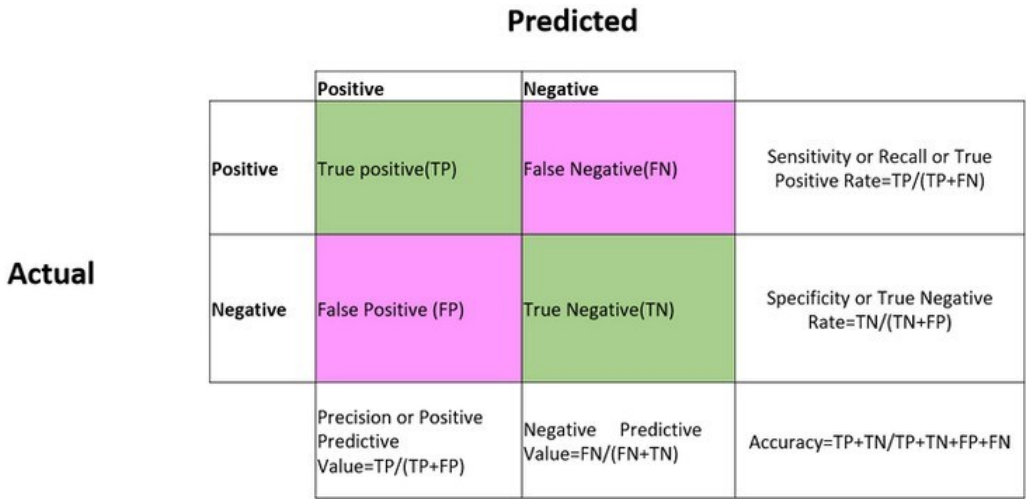
Com base nesses quatro conceitos iniciais são derivadas importantes métricas que são empregadas para avaliação dos modelos, quais sejam:

- *Accuracy*: a acurácia (ou Acc) é considerada uma das métricas mais simples e importantes. Ela avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas: $\lfloor (VP + VN)/(VP + VN + FP + FN) \rfloor$.
- Precisão ou *Precision* ou *Positive Predictive Value*: a precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos. O valor é obtido pela equação a seguir: $VP/(VP + FP)$.
- Sensibilidade ou *Recall* ou *Sensitivity* ou *True Positive Rate* (TPR): a sensibilidade avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. É o valor obtido pela equação: $VP/(VP + FN)$.
- *F1-score*: a *F-measure*, *F-score* ou *score F1* é uma média harmônica calculada com base na precisão e na sensibilidade. O valor é obtido da divisão de: $2\lfloor (Precision * Sensitivity)/(Precision + Sensitivity) \rfloor$.
- *Specificity* ou *True Negative Rate* (TNR): a especificidade avalia a capacidade do método de detectar resultados negativos. O valor é obtido pela equação: $VN/(VN + FP)$.
- *Receiver Operating Characteristic Curve* (ROC): a curva ROC, ou na tradução “Curva Característica de Operação do Receptor”, é um gráfico

que permite avaliar um classificador binário. Essa visualização leva em consideração a taxa de verdadeiros positivos (TVP; ou sensibilidade) e a Taxa de Falsos Positivos (TFP; ou 1 – especificidade). Essas taxas também podem ser referidas pelas siglas TPR (*True Positive Rate*) e FPR (*False Positive Rate*), respectivamente. Esse gráfico permite comparar diferentes classificadores e definir qual o melhor com base em diferentes pontos de corte. Na prática, quanto mais próximo do topo do eixo Y melhor o classificador. A curva é o gráfico formado pela interseção dos valores de S no eixo X e *Sensitivity* no eixo Y e o valor pode variar entre 0 e 1. A curva é gerada aplicando o detector várias vezes sobre o conjunto de teste enquanto varia-se algum parâmetro que afete precisão e sensibilidade.

- *Area Under the ROC Curve* (AUC): AUC é a métrica utilizada para avaliar a área sob a curva ROC (*Receiver Operating Characteristic*) sendo um valor escalar único que mede o desempenho geral de um classificador binário (Hanley; McNeil, 1982; 1983). O valor de AUC está dentro do intervalo (0,5-1,0), onde o valor mínimo representa o desempenho de um classificador aleatório e o valor máximo corresponderia a um classificador perfeito, com uma taxa de erro de classificação equivalente a zero.
- Matriz de Confusão: É uma maneira simples para representar os resultados de um modelo de classificação de dados. A Figura 1 traz uma demonstração da matriz de confusão.

Figura 1 – Métricas utilizadas em *DL*



Fonte: Kapoor, 2021.

A matriz de confusão indica a quantidade de ocorrências que o programa teve para cada uma das quatro categorias, quais sejam: VP, VN, FP e FN.

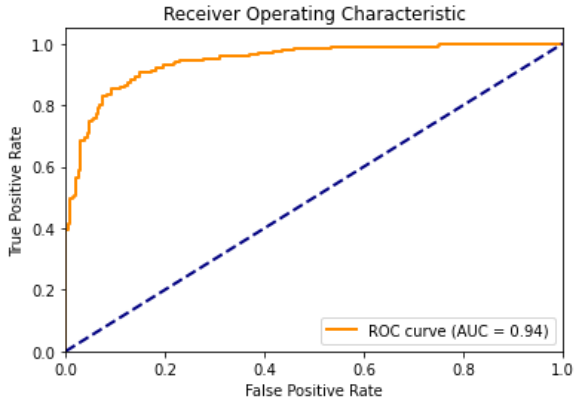
As métricas que serão utilizadas no presente trabalho, para efetuar as avaliações dos modelos, são a acurácia média ponderada (Acc), precisão (Prc), sensibilidade (Rec), F1-score (F1S) e AUC. Para subsidiar o entendimento, graficamente, também geramos a matriz de confusão e a curva ROC para o modelo implementado.

2.1.2 Métricas Mais Importantes de *Deep Learning* para o Contexto do Diagnóstico Médico

Em Ling, Huang e Zhang (2003), os autores discutem a utilização da precisão preditiva (acurácia) como critério principal e, muitas vezes, único para a avaliação de desempenho de algoritmos de aprendizado de classificação. Os autores provam formalmente que a AUC é uma métrica superior a acurácia. Eles apresentam definições rigorosas de consistência e discriminância para comparar duas medidas de avaliação de algoritmos de aprendizado, e através de avaliações empíricas e uma prova formal, estabelecem que a AUC é estatisticamente consistente e mais discriminante do que a precisão. Esse resultado é significativo, pois foi a primeira vez que se provou formalmente que a AUC é uma métrica melhor do que a acurácia na avaliação de algoritmos de aprendizado.

A constatação de Ling, Huang e Zhang (2003) tem implicações importantes para a avaliação, comparação e desenvolvimento de algoritmos de aprendizado profundo para a área médica. Isso porque a AUC-ROC é formada pela sensibilidade no eixo x e a especificidade no eixo y, conforme Gráfico 1, sendo essas duas métricas extremamente relevantes para o diagnóstico médico efetivo.

Gráfico 1 – AUC-ROC para um modelo DL



Fonte: o autor (2023).

Cabe, ainda, destacar que a AUC-ROC é uma métrica de avaliação importante para modelos de aprendizado profundo aplicados ao diagnóstico de imagens médicas pelas seguintes razões:

- desempenho geral do modelo: a AUC-ROC fornece uma única medida que resume o desempenho geral do modelo ao longo de vários limiares de classificação. Isso permite uma avaliação mais completa do modelo em comparação com métricas que dependem de um único ponto de corte, como precisão, sensibilidade ou especificidade;
- equilíbrio entre sensibilidade e especificidade: a curva ROC é construída plotando a taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos (1 – especificidade) para vários limiares de classificação. Isso permite visualizar o *trade-off* entre a sensibilidade e a especificidade do modelo e escolher o limiar que proporciona o melhor equilíbrio entre as duas;
- desempenho em diferentes limiares: ao contrário de outras métricas, como a precisão, a AUC-ROC leva em consideração o desempenho do modelo em todos os possíveis limiares de classificação, fornecendo uma visão mais completa de seu desempenho;
- comparação de modelos: a AUC-ROC é útil para comparar diferentes modelos, uma vez que proporciona uma medida única de desempenho. Um modelo com uma AUC-ROC maior é geralmente considerado melhor do que um modelo com uma AUC-ROC menor.

Portanto, a AUC-ROC é uma métrica valiosa para avaliar e comparar o desempenho de modelos de aprendizado profundo aplicados ao diagnóstico de imagens médicas.

Cabe ressaltar que a sensibilidade, também conhecida como taxa de verdadeiros positivos, é uma métrica crucial para a avaliação de modelos de aprendizado profundo em diagnósticos de imagens médicas devido às seguintes razões:

- identificação de condições médicas: em muitos contextos médicos, especialmente na detecção de doenças, é extremamente importante que um modelo de aprendizado profundo seja capaz de identificar corretamente todos os casos positivos (ou seja, condições patológicas). Uma alta sensibilidade indica que o modelo tem uma boa capacidade para detectar corretamente os verdadeiros positivos, reduzindo assim o risco de falsos negativos;
- consequências de FN: falsos negativos em diagnósticos médicos podem ter consequências graves, incluindo atrasos no tratamento e piora do prognóstico. Aumentar a sensibilidade minimiza a probabilidade de falsos negativos;
- equilíbrio com especificidade: a sensibilidade não deve ser considerada isoladamente, mas em conjunto com a especificidade (a capacidade de identificar corretamente os casos negativos). Embora seja importante maximizar a sensibilidade, é essencial garantir que isso não seja feito à custa de uma alta taxa de falsos positivos, que também podem ter consequências negativas, como tratamentos desnecessários ou estresse ao paciente.

Esses pontos reforçam que uma AUC-ROC elevada é um excelente indicador de capacidade de inferência de um modelo de classificação.

2.1.3 Conceitos Importantes Aplicáveis aos Modelos de VC

No contexto do estudo dos modelos de VC é importante entender os seguintes conceitos:

- *batch normalization*: segundo Goodfellow, Bengio e Courville (2016), a normalização em lote é uma técnica que fornece uma maneira elegante

de parametrizar novamente quase todas as redes profundas. Em termos mais gerais, essa é uma técnica de pré-processamento utilizada em redes neurais profundas para normalizar as entradas de cada camada da rede. O objetivo é acelerar o treinamento e melhorar a capacidade de generalização do modelo. O processo de normalização em lote consiste em calcular a média e o desvio padrão das ativações de uma determinada camada da rede em um conjunto de dados de treinamento, e em seguida, normalizar essas ativações usando a média e o desvio padrão calculados. A normalização é realizada de forma independente para cada exemplo de treinamento em um lote (daí o nome "*batch*"). A normalização em lote tem vários benefícios para o treinamento de redes neurais profundas, incluindo: redução da covariância entre as ativações das diferentes unidades da rede, o que pode levar a um treinamento mais rápido e a uma melhor generalização; redução do impacto de gradientes muito grandes ou muito pequenos, o que pode ajudar a evitar problemas de explosão ou desaparecimento de gradientes; e permitir que a rede treine com taxas de aprendizado mais altas, o que pode acelerar o treinamento e melhorar a precisão. Por fim, cabe destacar que a normalização em lote é amplamente utilizada em muitas arquiteturas de rede neural, incluindo redes convolucionais e redes totalmente conectadas, e geralmente é implementada como uma camada adicional na arquitetura da própria rede.

- *Cross-Dataset Evaluation* (CDE): a avaliação de conjuntos de dados cruzados refere-se à prática de testar um modelo de aprendizado de máquina em um conjunto de dados diferente daquele em que foi treinado. Em outras palavras, envolve avaliar o desempenho de um modelo em um conjunto de dados desconhecido, que pode ter diferentes características e distribuição de classes em comparação com o conjunto de dados de treinamento. A avaliação de conjuntos de dados cruzados é importante para verificar se um modelo é capaz de generalizar bem e lidar com a variabilidade presente em diferentes conjuntos de dados. Ela ajuda a avaliar a robustez do modelo e sua capacidade de realizar previsões precisas em situações do mundo real. Ao realizar a avaliação de conjuntos de dados cruzados, é importante garantir que o conjunto de dados de teste seja representativo o suficiente e cubra uma variedade de casos e condições.

Isso ajuda a identificar possíveis desafios e limitações do modelo, bem como oportunidades de melhoria. A avaliação de conjuntos de dados cruzados pode ser realizada de diferentes maneiras, como dividir o conjunto de dados em treinamento e teste usando diferentes proporções, utilizar conjuntos de dados completamente separados ou até mesmo realizar avaliações em conjuntos de dados públicos ou padrões estabelecidos pela comunidade. Sendo assim, a CDE é uma etapa crucial na validação e no teste de modelos de aprendizado de máquina, permitindo uma análise mais completa do desempenho e da capacidade de generalização do modelo em diferentes contextos.

- *data image augmentation*: segundo Shorten e Khoshgoftaar (2019) é uma técnica desenvolvida para reduzir o sobreajuste (*overfitting*), que ocorre quando o modelo “decora” os dados de treinamento, mas não obtém bons resultados com os dados de teste. Nos modelos de CNN, alguns parâmetros que podem ser ajustados são: *rotation, vertical shift, horizontal shift, zoom range, shear* e *crop* (Shorten; Khoshgoftaar, 2019). No entanto, de forma mais prática, *data augmentation* é usado basicamente para ampliar um conjunto de dados, adicionando variedade. Mas precisa-se tomar alguns cuidados. É preciso ter em mente que, embora ajude, não é uma solução para um conjunto de dados pouco representativo, porque mesmo com modificações, o *dataset* ainda estará limitado ao conjunto de dados original. Além disso, outro detalhe muito importante é que a técnica só deve ser aplicada ao conjunto de treino. Outro ponto a ser considerado é se as modificações realizadas nas imagens são condizentes com as variações que serão de fato observadas nas imagens contidas nos *datasets* que serão utilizados em ambiente de produção.
- *dropout*: segundo Srivastava *et al.* (2014), é uma técnica de regularização que zera os valores de ativação de neurônios escolhidos aleatoriamente durante o treinamento. Essa restrição força a rede a aprender recursos mais robustos, em vez de depender da capacidade preditiva de um pequeno subconjunto de neurônios na rede. Esse conceito foi estendido para CNN com *spatial dropout* (Tompson *et al.*, 2015), que exclui mapas de características inteiras em vez de neurônios individuais.

- *ensemble*: um "*ensemble*" é uma abordagem no aprendizado de máquina que combina previsões de múltiplos modelos para produzir uma previsão ou classificação mais precisa e robusta, visando aprimorar a precisão e a robustez da previsão. Esse método capitaliza na ideia de que a união de diferentes modelos ou hipóteses individuais forma uma hipótese mais forte, especialmente quando esses modelos possuem diferentes pontos fortes e susceptibilidades a erros. As técnicas de *ensemble*, como *bagging* (ilustrada pelo algoritmo *Random Forest*), *boosting* (como o *Gradient Boosted Trees* e *AdaBoost*) e *stacking* (que combina previsões usando um modelo meta), são essenciais para reduzir *overfitting*, amplificar a precisão e conferir maior robustez ao modelo final, protegendo-o contra falhas específicas de modelos singulares.
- *Grad-CAM* (*Gradient-Weighted Class Activation Mapping*) é uma técnica que oferece visualizações das regiões de uma imagem que influenciaram a decisão de modelos de aprendizado profundo, como CNN. Utilizando gradientes da saída da classe-alvo em relação a uma camada específica da CNN, cria-se um mapa de calor que destaca áreas cruciais para a decisão da rede. Esta abordagem promove interpretabilidade, ajuda a identificar decisões baseadas em regiões irrelevantes da imagem e valida se modelos pré-treinados estão focalizando corretamente ao serem adaptados para novas tarefas. O *Grad-CAM* é particularmente valioso em aplicações como diagnósticos médicos, onde a compreensão da decisão do modelo é essencial.
- *Gradient Vanish*: o desaparecimento do gradiente (em tradução livre para o português) é um problema que ocorre durante o treinamento de redes neurais artificiais usando algoritmos de aprendizado baseados em gradiente, como a retropropagação (*backpropagation*). O problema foi tratado em Pascanu, Mikolov e Bengio (2013). Este problema torna-se particularmente pronunciado em redes neurais muito profundas, frequentemente chamadas de DL. Durante o treinamento de uma rede neural, os pesos são ajustados em proporção ao gradiente da função de perda com respeito aos pesos atuais. Se esse gradiente se tornar muito pequeno, os pesos da rede neural não mudam muito e, consequentemente, a rede para de aprender. Esse é o problema do desaparecimento do gradiente.

Quando se utiliza a função de ativação sigmoide ou a tangente hiperbólica, os gradientes podem se tornar muito pequenos se os valores de entrada para essas funções estiverem muito distantes de zero. À medida que esses pequenos gradientes são retropropagados na rede durante o treinamento, eles podem ficar cada vez menores, especialmente em redes muito profundas, até que tenham pouco ou nenhum efeito na atualização dos pesos — e é por isso que chamamos este problema de "desaparecimento do gradiente". Existem várias técnicas propostas para aliviar o problema do desaparecimento do gradiente, incluindo o uso de funções de ativação alternativas (como ReLU), inicialização de pesos cuidadosa [como a inicialização de He *et al.* (2021) ou Glorot e Bengio (2010)], normalização em lotes (*batch normalization*), e o uso de arquiteturas de rede que permitem o aprendizado de representações em diferentes níveis de abstração (como redes residuais).

- *transfer learning*: segundo Goodfellow, Bengio e Courville (2016), o aprendizado por transferência é uma técnica de treinamento de modelos de aprendizado de máquina que consiste em aproveitar o conhecimento prévio de um modelo treinado em uma tarefa relacionada para auxiliar o treinamento de um novo modelo em uma tarefa diferente, possivelmente com um conjunto de dados menor. A ideia principal do aprendizado por transferência é que o conhecimento adquirido pelo modelo treinado em uma tarefa relacionada pode ser útil para auxiliar o treinamento de um novo modelo em outra tarefa. Por exemplo, se um modelo foi treinado para reconhecer imagens de carros em um grande conjunto de dados, o conhecimento adquirido pode ser transferido para ajudar no treinamento de um modelo para reconhecer motocicletas em um conjunto de dados menor. O aprendizado por transferência pode ser aplicado de várias maneiras, como utilizar uma rede pré-treinada como uma “rede base” e ajustar os pesos da última camada para a nova tarefa, ou utilizar a rede pré-treinada como um “extrator de características” para extrair representações úteis do conjunto de dados original que podem ser utilizadas para treinar um novo modelo. O aprendizado por transferência é útil porque pode ajudar a melhorar a precisão do modelo, especialmente quando há poucos dados disponível para a nova tarefa. Além disso, ele pode ajudar

a reduzir o tempo e o esforço necessários para treinar um novo modelo, já que parte do conhecimento necessário já foi adquirido pelo modelo pré-treinado. Em termos gerais, normalmente, se aplica *transfer learning* utilizando alguma arquitetura pré-treinada em um dataset genérico e grande, como o *ImageNet*, e depois se aplica esse modelo um *dataset* menor e de problema específico. A premissa é que as características extraídas das imagens para um *dataset* ou problema genérico e grande são também genéricas o suficiente para servirem para outros domínios de problema mais específicos.

2.2 LEVANTAMENTO BIBLIOGRÁFICO SOBRE APLICAÇÃO DE VISÃO COMPUTACIONAL EM IMAGENS RADIOGRÁFICAS

A revisão da literatura foi utilizada para alcançar o objetivo de aprofundar o conhecimento do *status* da pesquisa atual sobre a aplicação de DL no domínio de imagens médicas, mais especificamente imagens de radiografia de tórax com a intenção de classificar imagens contendo pulmões saudáveis ou infectados por COVID-19.

Para alcançar esses objetivos seguiu-se o paradigma *Goal-Question-Metric* (GQM) proposto por Basili, Caldiera e Rombach (1994).

2.2.1 Metodologia do Levantamento Bibliográfico

2.2.1.1 Planejamento

Dispusemos os objetivos específicos em dois grupos, a saber:

- mapear os modelos e métricas publicados para identificar e classificar quais foram os modelos que se mostraram mais eficazes, e quais as métricas obtidas em cada um deles na tarefa de classificação de imagens radiográficas (Raio X e TC) com a finalidade de detectar pulmões saudáveis ou com COVID-19.
- mapear quais especificidades impactaram os resultados alcançados pelos modelos e que foram reportadas nos artigos. Nessa linha, a pesquisa está atenta ao emprego de técnicas como *transfer learning, data aug-*

mentation e outros detalhes como tamanho das imagens, e tamanho, disponibilidade e vieses dos *datasets* utilizados nos estudos.

2.2.1.2 Condução

Para condução do trabalho foi empregado o protocolo aconselhado para a Revisão Sistemática da Literatura, seguindo o seguinte:

a. Definição das questões de pesquisa:

Foi definida uma questão principal (Q1) e duas subquestões conforme consta abaixo:

Questão principal de pesquisa:

Q1: Quais modelos, métodos e técnicas de classificação de imagens são eficazes quando aplicados em imagens radiográficas de tórax com a intenção de classificar imagens contendo pulmões saudáveis ou infectados por COVID-19?

Subquestão de pesquisa 1 – Modelos e métricas:

SQ1: Quais foram os modelos que se mostram mais eficazes, e quais as métricas obtidas para cada um deles?

Subquestão de pesquisa 2 – Especificidades que impactaram os resultados:

SQ2: Quais foram as especificidades do emprego de técnicas como *transfer learning, data augmentation* e outros detalhes como tamanho das imagens, e tamanho, disponibilidade e vieses dos *datasets* utilizados nos estudos?

b. Estratégia de Busca:

A estratégia de busca automática foi aplicada nas bases *Scopus* e *Web of Science* utilizando o *proxy* autenticado da Universidade Tecnológica Federal do Paraná (UTFPR). Nessa etapa, foi delimitada a janela de publicação dos *papers* para o período de 2020 a 2023.

c. *String* de Busca:

A *string* de busca utilizada foi a seguinte:

(TITLE-ABS-KEY ("*chest x-ray*") AND (CT) AND (“*classification”*) AND (CNN) OR (“*Vision Transformer*”) OR (“*Swin Transformer*”) AND (COVID)) AND PUBYEAR > 2019

2.2.1.3 Aplicação dos critérios de exclusão

2.2.1.3.1 Critérios de Seleção (Critérios de Exclusão)

O primeiro critério de exclusão foi retirar da pesquisa aqueles artigos que não estavam redigidos na língua inglesa. Um conjunto de termos foi levantado, a partir da leitura dos títulos e das palavras-chave e foram utilizados como:

Critérios de Exclusão, conforme a Quadro 1:

Quadro 1 – Critérios de exclusão de artigos

Brain	*Cardiovascular*	*Nodules*	*Segmentation*
Breast	*Clavicle And Rib*	*Pelvic*	*TB*
Cancer	*Detection*	*Pneumoconiosis*	*Tong*
Carcinoma	*Ensemble*	*Pneumothorax*	*Toracic Trauma*
Cardiac	*Gastric*	*Random Forest*	*Tuberculosis*
Cardiomegalia	*Heart Disease*	*Renal*	*Tumor*

Fonte: o autor (2023).

2.2.1.4 Seleção e análise dos trabalhos

a) Resultados da primeira busca:

A pesquisa inicial resultou em 1.430 artigos. Utilizando o *software Zotero*, foram extraídos 445 desses artigos que estavam disponíveis em formato PDF a partir dos arquivos "ris" gerados nas bases de dados.

Seguindo os critérios de exclusão, procedemos com a deduplicação e selecionamos apenas os artigos que correspondiam aos critérios de interesse, levando em consideração seus títulos e resumos. Esse processo resultou em 79 artigos.

Aplicamos critérios adicionais de seleção, incluindo o ano de publicação (2021, 2022 e 2023) e o foco temático em modelos de CNN, ViT e *Swin Transformer*. Também levamos em consideração a classificação Qualis dos artigos: para os artigos sobre CNN, selecionamos apenas aqueles classificados como A1 ou A2. Já para os artigos sobre ViT e *Swin Transformer*, ampliamos a seleção para incluir classificações A1, A2, A3 ou A4.

No final deste processo de seleção, restaram 36 artigos sendo: 18 sobre CNN, 13 sobre ViT e cinco sobre *Swin Transformer*. Esses artigos, todos primários, compõem a base da nossa revisão de literatura sobre os modelos.

Procedemos então à leitura mais detalhada das seções de resultados e conclusões desses 36 artigos selecionados para aprofundar nossa compreensão e análise.

2.2.1.5 Artigos complementares ao levantamento bibliográfico

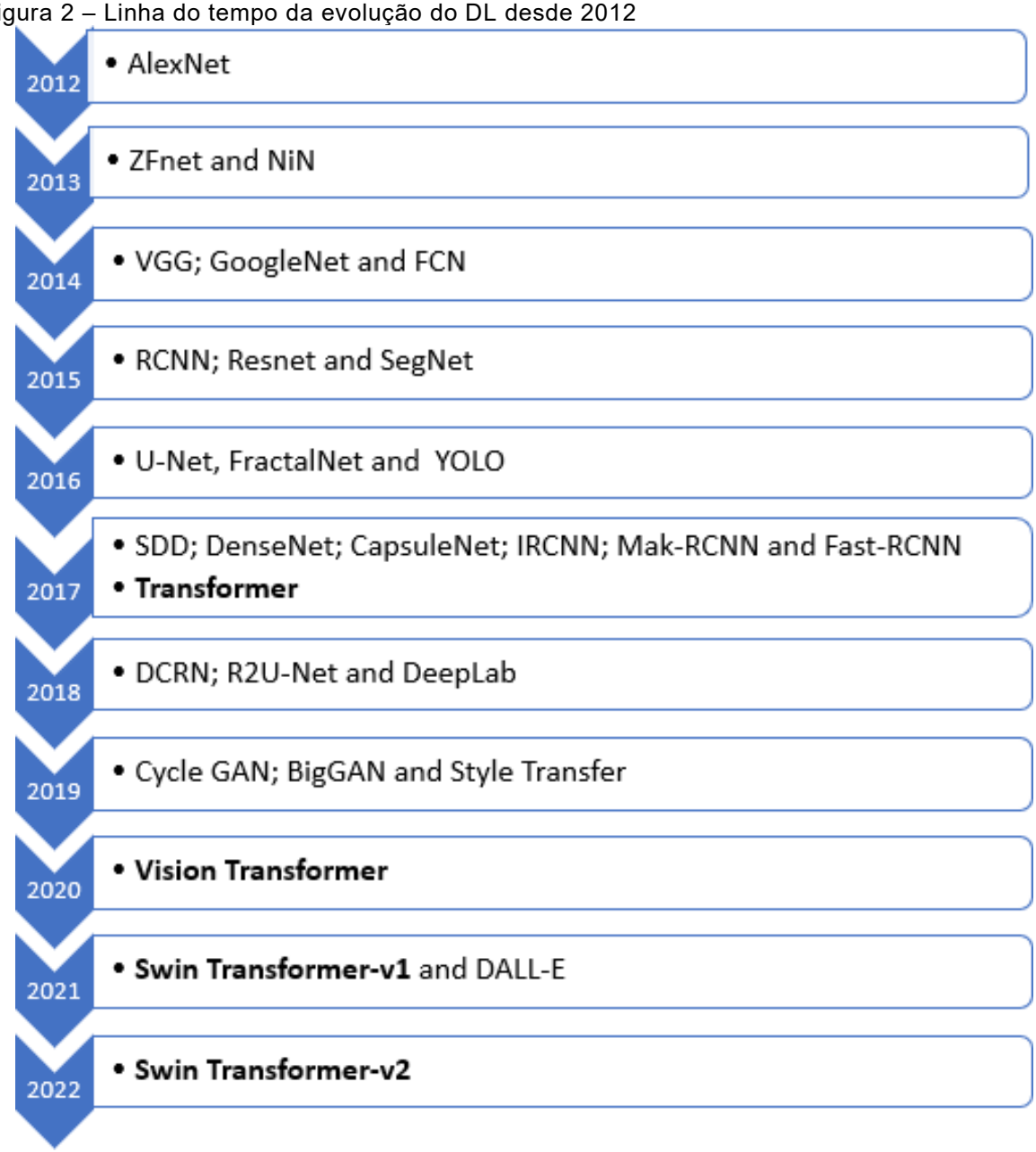
Além dos 36 artigos primários, foram escolhidos mais de 50 artigos contendo revisões sistemáticas, citações sobre técnicas e assuntos específicos e pertinentes para dar o embasamento necessário ao presente trabalho. Além dos artigos, também estão sendo citados conteúdos referentes a quatro livros importantes para a área de estudo e alguns sites.

2.2.2 Visão Geral Sobre a Evolução dos Modelos de VC

A área de VC tem experimentado um progresso notável com a aplicação de técnicas de aprendizado profundo, especialmente com a adoção de CNN e mais recentemente com os ViT.

No levantamento bibliográfico que foi realizado foi possível estabelecer uma linha do tempo sobre a evolução dos modelos de VC. A esquematização desse levantamento encontra-se demonstrada na Figura 2.

Figura 2 – Linha do tempo da evolução do DL desde 2012



Fonte: o autor (2023).

Para facilitar o entendimento separamos cada uma das arquiteturas, que serão abordadas no presente trabalho, em subseções que serão aprofundadas abaixo.

2.2.2.1 Modelo *ResNet*

O estudo da VC aplicada a imagens médicas, como expõe Liu *et al.* (2017), presenciou avanços consideráveis na última década. Esses progressos foram possíveis graças aos notáveis avanços da área de VC a partir de 2012,

com as primeiras publicações sobre a utilização de CNN emergindo nesse mesmo ano.

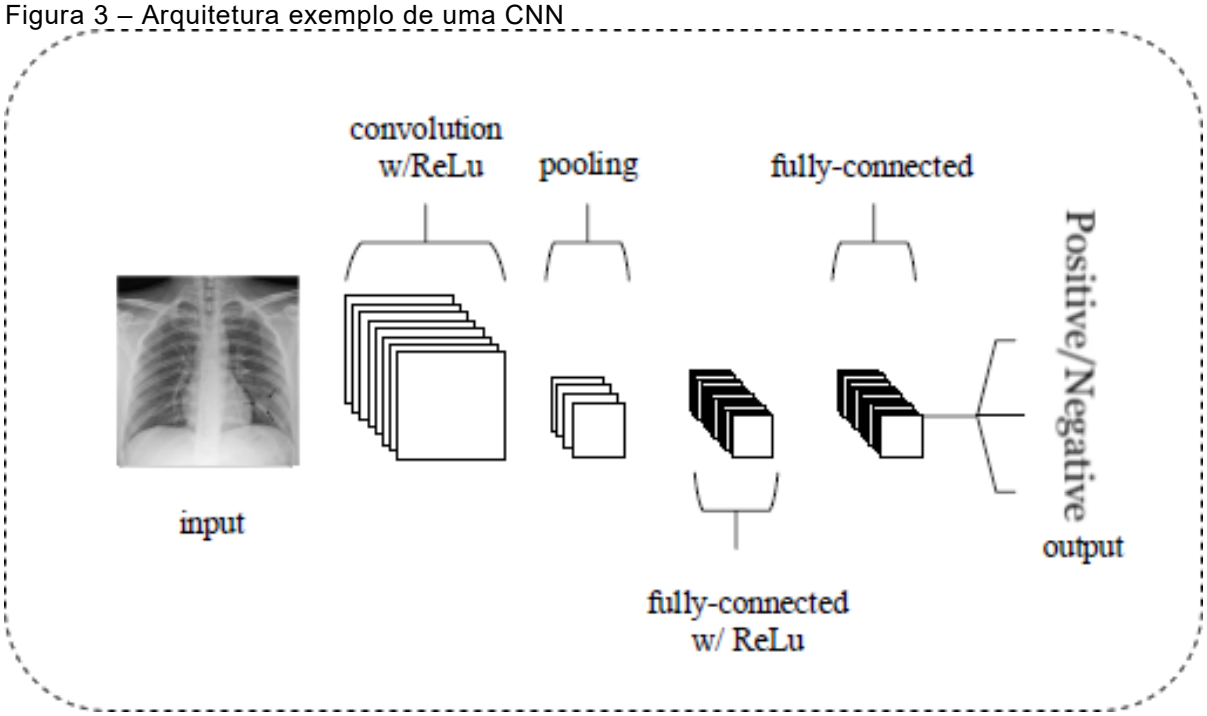
Antes de 2012, o desafio *ImageNet Large Scale Visual Recognition Challenge (*ILSVRC*)*, que desempenha um papel fundamental no avanço da VC, era dominado por abordagens de ML que exigiam a seleção manual de características para ajustar os modelos. No entanto, um trabalho de Krizhevsky, Sutskever e Hinton (2012) superou significativamente os melhores resultados de modelos anteriores, ao introduzir uma arquitetura de CNN. Esse trabalho é amplamente reconhecido como um marco importante na área tendo sido apresentado no ILSVRC 2012.

Desde então, surgiram diversos estudos empregando modelos baseados em CNN. Alguns dos modelos mais notáveis incluem VGG-16 (Geng *et al.*, 2019), *Inception-v3* (Szegedy *et al.*, 2016), Redes Residuais 1 e 2 (ResNet1 e ResNet2) (Jung; Chi, 2020; He *et al.*, 2016), Redes de Convoluções Separáveis em Profundidade (Xception) (Chollet, 2017), Redes Densamente Conectadas (DenseNet) (Yao *et al.*, 2020), entre outros. Esses modelos baseados em CNN são frequentemente utilizados para implementar sistemas para tarefas de VC aplicáveis a *Computer-Aided Detection* (CADe) e *Computer-Aided Diagnosis* (CADx) (Bakator; Radosav, 2018).

Um desafio para a área de VC é a demanda por alto poder computacional para as tarefas de VC, como discutido por Rawat e Wang (2017). Apesar das CNN terem sido apresentadas em 1998 por Lecun *et al.* (1998), foi apenas recentemente que os recursos computacionais foram capazes de suportá-las em larga escala, em grande parte impulsionados pelo avanço das *Graphic Processing Units* (GPU) a partir de 2007 (Feng *et al.*, 2019). A relação custo-benefício se tornou vantajosa com o lançamento da série *GeForce 10* da *Nvidia*, em 2016. Vale ressaltar que a evolução dos modelos de DL é contínua, com a comunidade científica constantemente produzindo novos modelos e modificando os existentes.

Uma CNN utiliza a técnica *feedforward* para analisar imagens através do processamento de dados com topologia semelhante a um *grid*, pode ser empregada para detectar e classificar objetos em imagens, como exemplificado na Figura 3, que apresenta uma representação simplificada da arquitetura de uma CNN para classificar imagens de Raio X com COVID-19.

Figura 3 – Arquitetura exemplo de uma CNN

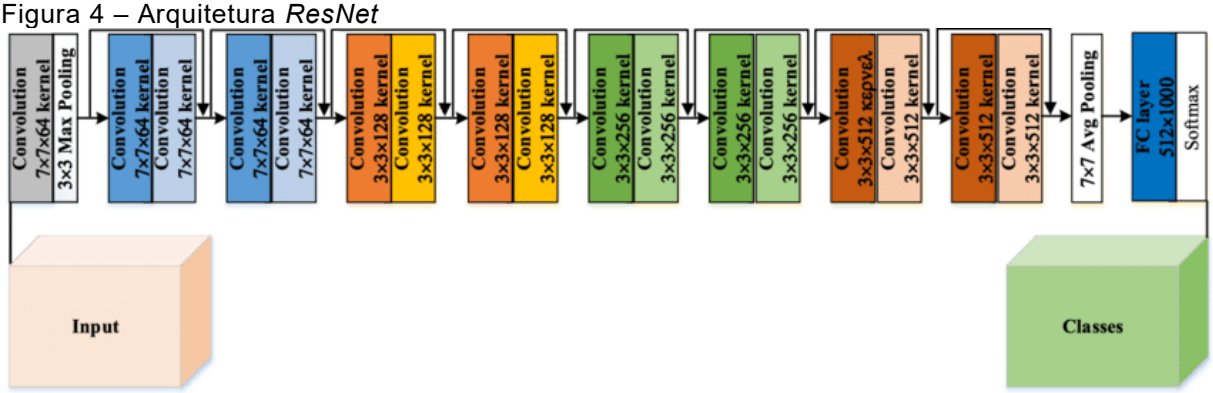


Fonte: Rafi (2020).

No entanto, à medida que as redes ficavam mais profundas, elas começavam a enfrentar problemas de degradação. He *et al.* (2016) abordaram esse problema com a introdução do modelo *Residual Network* (*ResNet*), uma arquitetura que incorporava conexões residuais para facilitar a propagação de gradientes através de redes muito profundas. Os principais pontos abordados no artigo foram:

- arquitetura *ResNet*: a arquitetura *ResNet* introduzida por He *et al.* (2016) é caracterizada por "blocos residuais", que contêm atalhos ou conexões de salto que permitem que o gradiente seja retropropagado para camadas anteriores. Estes atalhos ou conexões residuais ajudam a aliviar o problema do desaparecimento do gradiente, permitindo que redes muito profundas sejam treinadas efetivamente. O elemento-chave de um bloco residual é a "conexão de atalho" que pula uma ou mais camadas. A Figura 4 demonstra a arquitetura da rede *ResNet*.

Figura 4 – Arquitetura *ResNet*



Fonte: He *et al.* (2016).

- desempenho em profundidade: As redes residuais demonstraram que as redes podem de fato se beneficiar de serem mais profundas. He *et al.* (2016) treinaram *ResNet* de até 152 camadas, enquanto ainda obtinham uma melhoria no desempenho. Eles conseguiram isso sem aumentar a complexidade computacional em comparação com as redes menos profundas.
- resultados experimentais: o artigo relata resultados experimentais obtidos em várias competições de *benchmark* e desafios de reconhecimento de imagens. As *ResNet* alcançaram resultados recordes em vários desses desafios, demonstrando a eficácia de sua arquitetura. No desafio da ImageNet, por exemplo, uma *ResNet* de 152 camadas venceu a competição com um erro de top-5 de 3,57%, que foi um recorde na época.

Esse artigo trouxe um impacto significativo para a área de aprendizado profundo e redes neurais convolucionais, visto que a arquitetura *ResNet* se tornou uma das principais escolhas para muitas tarefas de VC. Ele também abriu o caminho para redes ainda mais profundas e outras arquiteturas que utilizam conexões de atalho.

Chen *et al.* (2021) propuseram uma revisão sistemática sobre a evolução e aplicação das CNN na classificação de imagens. O estudo enfatiza tendências atuais, como a importância dos mecanismos de atenção, o design de redes otimizado para plataformas móveis, a escolha estratégica de hiperparâmetros, a aplicação de aprendizado por transferência e técnicas de aumento de dados, bem como a relevância das estratégias de treinamento. Ademais, a pesquisa destaca desafios inerentes aos modelos de CNN, tais como o equilíbrio entre precisão e eficiência, e a transição para aprendizado semi-supervisionado ou

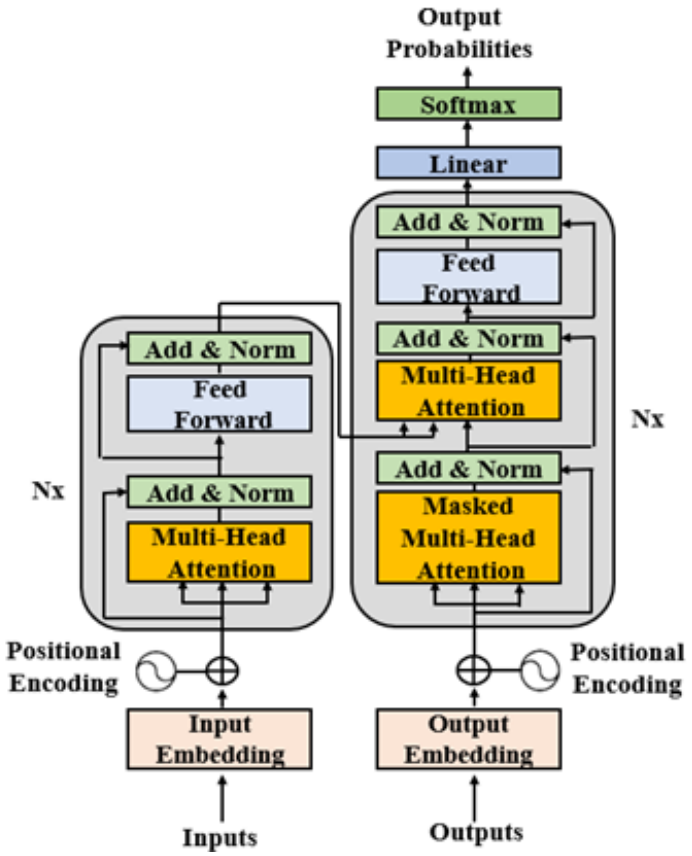
não supervisionado. A revisão também sugere direções futuras de pesquisa, incluindo a integração eficaz de convolução e *Transformer*, e a necessidade de investigação adicional sobre componentes convencionais nas CNN. Portanto, o artigo de Chen *et al.* (2021) fornece um panorama abrangente e perspicaz das atuais tendências, desafios e futuras direções na área de classificação de imagens utilizando CNN.

2.2.2.2 Modelo *Transformer*

Em paralelo ao desenvolvimento das CNN, uma nova abordagem para o processamento de sequências foi introduzida por Vaswani *et al*. (2017), denominada *Transformer*. Embora originalmente concebido para o PLN, o *Transformer* logo também mostrou grande potencial para tarefas de VC pelo fato desta arquitetura inovadora ser baseada inteiramente no mecanismo de atenção, eliminando a necessidade de redes recorrentes ou convoluções. Os principais pontos abordados no artigo foram:

- mecanismo de atenção: a principal inovação do *Transformer* é o uso de "atenção autorregressiva", que calcula a importância relativa de cada palavra em relação a todas as outras na mesma frase. Isto permite que o modelo dê mais "atenção" para palavras relevantes para a tarefa atual, em vez de tratar todas as palavras igualmente.
- arquitetura *Transformer*: ao contrário das arquiteturas RNN e CNN, o *Transformer* é construído inteiramente a partir de camadas de atenção autorregressiva. Isto permite que ele capte dependências de longo alcance entre palavras sem a necessidade de recorrência ou convoluções. Além disso, os *Transformers* são altamente paralelizáveis, o que os torna mais eficientes para treinamento em *hardware* moderno. A Figura 5 demonstra a arquitetura da rede *Transformer*.

Figura 5 – Arquitetura *Transformer*



Fonte: Vaswani *et al*. (2017).

- resultados experimentais: o artigo apresentou resultados experimentais que mostram que os *Transformers* superam os modelos de SOTA em várias tarefas de PLN, incluindo a tradução automática. Eles também são mais eficientes em termos de computação, sendo capazes de treinar em menos tempo do que os modelos comparáveis.
- aprendizado de representações contextuais: ao contrário de modelos anteriores, o *Transformer* é capaz de aprender representações contextuais de palavras. Isto significa que ele pode aprender que a mesma palavra pode ter significados diferentes dependendo do contexto em que é usada.

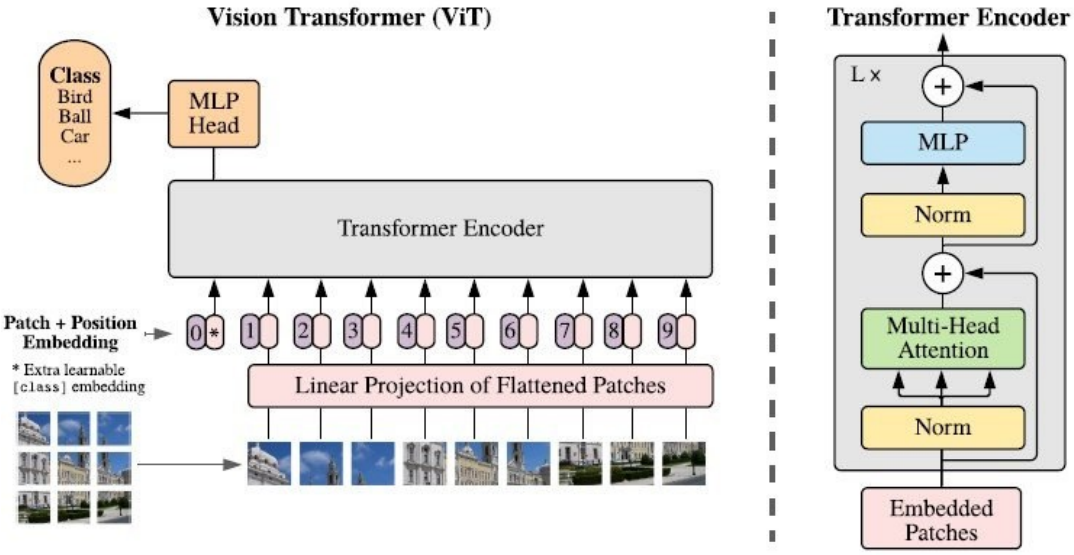
O artigo de Vaswani *et al*. (2017) tem tido um impacto significativo no campo do processamento de linguagem natural. A arquitetura *Transformer* serve de base para modelos posteriores, como o BERT, GPT e muitos outros, que são amplamente usados em uma variedade de tarefas de PLN.

2.2.2.3 Modelo *Vision Transformer* (ViT)

Dosovitskiy *et al*. (2021) propuseram o ViT, um modelo de aprendizado profundo que se baseia na arquitetura *Transformer* (tradicionalmente aplicada em PLN). Os autores mostraram que, quando treinados com dados suficientes, os ViT poderiam superar as CNN no desempenho de classificação de imagens. Os principais pontos abordados no artigo foram:

- *patch-based input*: o ViT divide uma imagem de entrada em *patches* fixos, como se fossem "palavras" em um texto, os quais são então achatados e transformados linearmente para criar uma sequência de *embeddings* de dimensão fixa.
- *positional encoding*: como os *Transformers* não têm consciência da estrutura espacial 2D das imagens, o ViT usa codificações posicionais, assim como os *Transformers* em NLP, para infundir informações sobre a localização relativa dos *patches* na imagem.
- arquitetura ViT: após a criação dos *embeddings* de *patch*, o ViT aplica várias camadas do *Transformer*, cada uma das quais consiste em atenção multicabeça (*Multi-head Self Attention layers* — MSA) e redes *feedforward*, para processar esses *embeddings*. Isso permite que o modelo capture interações entre todos os pares de *patches*. A Figura 6 demonstra a arquitetura da rede ViT.

Figura 6 – Arquitetura *Vision Transformer (ViT)*



Fonte: Dosovitskiy *et al.* (2021).

- escala de treinamento: o ViT alcança melhores resultados quando é treinado com grandes volumes de dados e quando é treinado por mais tempo. Portanto, embora o ViT possa superar as CNN em algumas tarefas, isso geralmente só é verdade quando há uma quantidade suficiente de dados disponíveis para o treinamento e a um custo de computação alto.
- transferibilidade: assim como os modelos baseados em *Transformers* em NLP, os modelos ViT treinados em grandes volumes de dados também apresentam desempenho promissor quando transferidos para outras tarefas com poucos dados de treinamento disponíveis.
- interpretabilidade: devido à natureza da atenção no modelo, o ViT é um pouco mais interpretável do que uma CNN. A atenção do modelo pode ser visualizada para entender onde o modelo está "olhando" ao tomar decisões.
- ausência de convolução: ao contrário das abordagens convencionais de VC, que usam extensivamente convoluções para processar imagens, o ViT não usa convoluções. Isso representa uma mudança significativa na forma como os modelos de VC são geralmente construídos.
- custo computacional caro para tarefas de grande escala: Os mecanismos tradicionais de autoatenção usados em *Transformers* têm uma complexidade computacional quadrática em relação ao comprimento da sequência de entrada, o que pode ser computacionalmente caro para tarefas em grande escala.

Segundo Sun *et al*. (2022), no modelo ViT, o módulo de autoatenção Att() aprende três matrizes de pesos. São elas: WQ, WK, WV. Com base nelas, x é projetado na consulta (Q), chave (K) e valor (V). Uma matriz de atenção A é geralmente calculada por uma função de similaridade S(·) sobre consultas e chaves. Na autoatenção padrão, S(·) é a normalização softmax. As saídas do módulo de autoatenção são, portanto, $O = Att(x) = AV$, onde $A \in R^{N*N}$ sofre de uma complexidade quadrática espaço-temporal em relação para o número de patch N. Portanto, a complexidade de computação teórica neste caso é $O(N^2 d) = O\lfloor(H^2 W^2/p^4)d\rfloor$. Consequentemente, o módulo de autoatenção torna-se sensível ao tamanho da imagem, sofrendo com o aumento da altura (H) e largura (W)

além do tamanho dos *patches* p. Assim, a complexidade quadrática do modelo ViT é, por definição, seu principal gargalo computacional.

Deve-se notar que a arquitetura ViT tende a ser computacionalmente mais cara que a CNN, quando aplicado a grandes volumes de dados ao mesmo tempo que tende a um desempenho melhor do que os modelos CNN (Dosovitskiy *et al*., 2020; Tuli; Dasgupta; Grant; Griffiths, 2021; Wei *et al*., 2022). Outra limitação do ViT é que sua complexidade computacional é quadrática ao tamanho da imagem, ou ao número de *patches*, tornando-o inadequado para imagens de alta resolução.

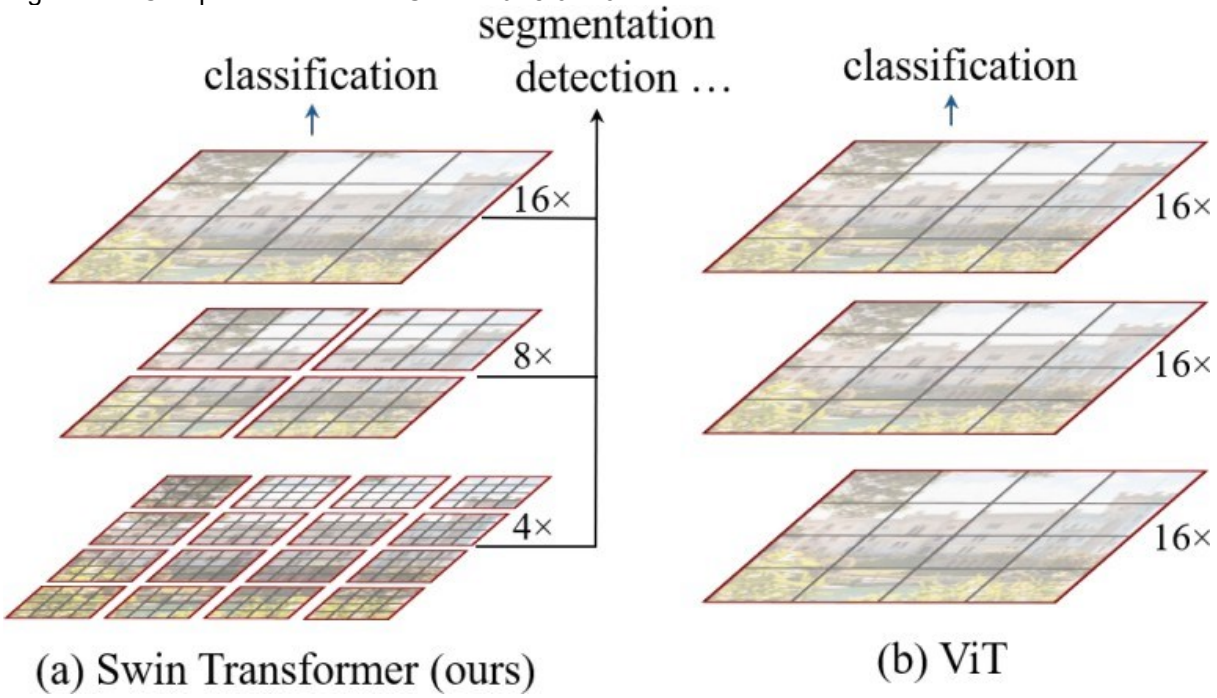
2.2.2.4 Modelo *Swin Transformer* V1

Baseado nos avanços do ViT, Liu, Shih e Zhong (2021) apresentaram o *Swin Transformer*, que implementa uma estrutura hierárquica e uma atenção localizada para tornar os *Transformers* mais eficientes para tarefas de visão.

No presente trabalho, o *Swin Transformer* foi empregado aplicando transferência do aprendizado realizado no conjunto de dados *ImageNet-1K* (Liu *et al*. 2021). A palavra *Swin* (em *Swin Transformer*) tem origem em *Shifted Windows*. A técnica da janela deslocada permite que a rede aprenda atenção em diferentes escalas espaciais. Isso é alcançado através de uma hierarquia de camadas que dividem a entrada em resoluções cada vez menores, permitindo que cada camada aprenda atenção em diferentes escalas. Essa técnica combina a atenção aprendida em diferentes escalas para produzir uma representação final da entrada.

O conceito de janela deslocada não é novo para a comunidade de pesquisa. Ele tem sido usado em CNN há muitos anos. É um dos recursos da CNN que a destacou no campo da VC, pois trouxe grande eficiência. No entanto, ainda não havia sido utilizado em combinação com as estratégias do ViT. A Figura 7 mostra uma comparação entre a forma como os *patches* são segmentados nos modelos ViT e *Swin Transformer*.

Figura 7 – Comparativo entre o *Swin Transformer* e o *ViT*



Fonte: Dosovitskiy *et al*. (2020); Liu *et al*. (2021).

Semelhante ao modelo ViT, o *Swin Transformer* usa *patches*, porém, em vez de usar um tamanho fixo, como 16x16 *pixels*, o *Swin Transformer* começa com pequenos *patches* (4×4 *pixels*) na primeira camada do *Transformer* e vai mudando o tamanho dos *patches* nas camadas mais profundas. O modelo mescla essas camadas menores em camadas maiores à medida que se aprofunda na arquitetura de rede. Ele pega uma imagem e a divide em *patches* de 4×4 pixels. Cada *patch* é uma imagem colorida com três canais que podem estar no padrão Vermelho, Verde e Azul (RGB), por exemplo. Assim, um *patch* tem um total de 48 dimensões de recursos. Ou seja, 4×4×3 = 48.

Ele é então linearmente transformado em uma dimensionalidade chamada C, de sua escolha. Até este ponto, em comparação com o ViT, os *patches* de imagem são menores em tamanho. O valor, C, determina o tamanho do seu modelo *Transformer*.

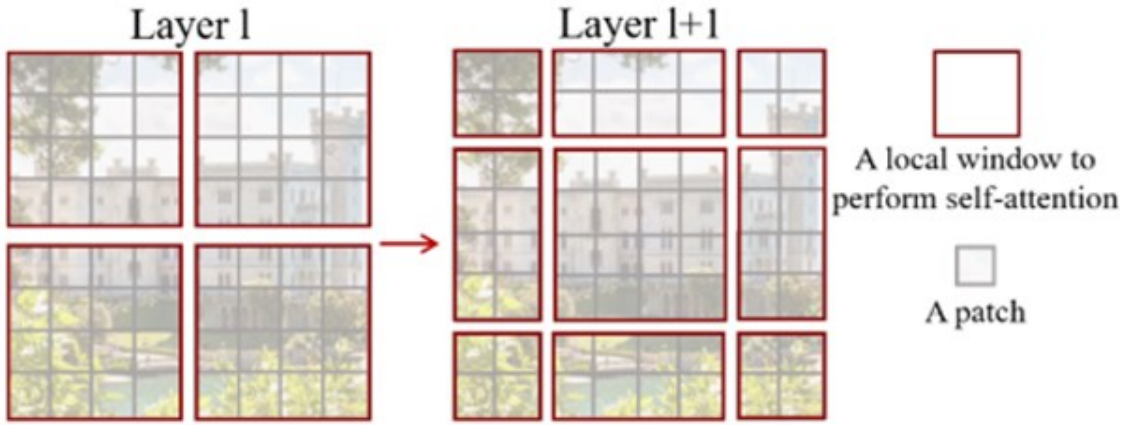
A abordagem empregada pelo *Swin Transformer* é conhecida como janela deslocada (LIU *et al*., 2021), e consiste em computar a autoatenção dentro de janelas locais, ao invés de computar dentro de um campo receptivo global como o ViT faz. Uma janela de deslocamento contém *patches* não sobrepostos de MxM (onde M = 7 é o tamanho da janela) e a autoatenção é calculada nessa janela.

Como resultado, o ViT MSA original, que tem uma complexidade computacional quadrática em relação ao número de *patches* e HxW, se torna muito menos computacionalmente caro no contexto do *Swin Transformer*. Isso ocorre porque a complexidade do MSA baseado na janela do *Swin Transformer* (W-MSA) é linear. A equação que representa a complexidade do ViT é $\Omega(MSA) = 4hwC^2 + 2(hw)^2C$, enquanto a equação para o *Swin Transformer* é $\Omega(W\ MSA) = 4hwC^2 + 2M^2hwC$. Essas equações ilustram que o custo computacional do *Swin Transformer* é substancialmente menor. Nas fórmulas, o MSA simboliza o mecanismo de autoatenção usado no ViT, enquanto o W-MSA simboliza o mesmo mecanismo do *Swin Transformer*.

Como pode ser visto, o modelo *Swin Transformer* adiciona uma complexidade computacional linear ao tamanho da imagem de entrada. Ele calcula a autoatenção apenas dentro da janela local e não globalmente, como no modelo ViT. Esse recurso permite que o modelo execute tarefas de reconhecimento densas e permite que ele seja usado para tarefas de VC de uso geral com imagens coloridas maiores.

No *Swin Transformer*, a saída de uma camada é mesclada por uma *Merging Layer*, que concatena os vetores de grupos de *patches* das vizinhanças na imagem cada vez que a janela de atenção muda em relação à camada anterior. Por exemplo, se na primeira camada a atenção foi limitada à vizinhança dessas regiões, na próxima camada as regiões são deslocadas (como na *strided convolution*). A Figura 8 mostra a abordagem da janela deslocada.

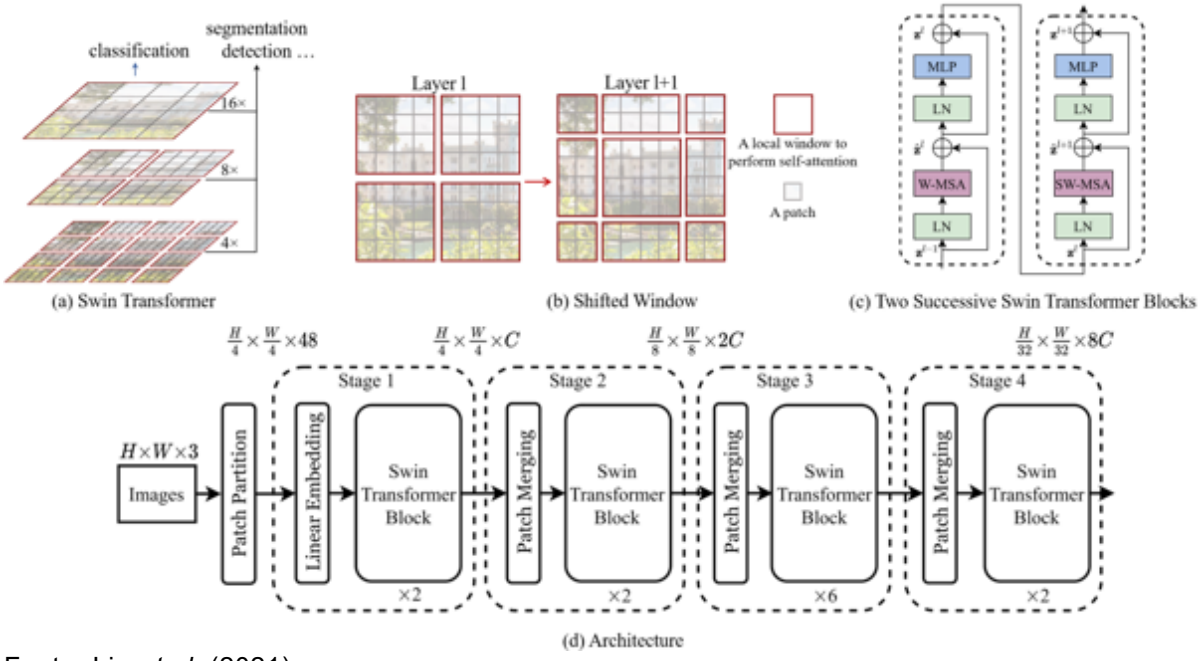
Figura 8 – Janela Deslocada do *Swin Transformer*



Fonte: Liu *et al*. (2021).

Os *patches* que chegaram em janelas separadas na primeira camada e não puderam se comunicar, portanto, podem fazê-lo na camada dois. Esses *patches* resultantes são mesclados pela camada de mesclagem. Esse processo é repetido dependendo do número de camadas escolhidas. Abaixo, a Figura 9 resume como o *Swin Transformer* funciona.

Figura 9 – (a) Estratégia de divisão de *patches* do *Swin Transformer;* (b) Abordagem de janela deslocada; (c) Blocos do *Swin Transformer;* (d) Visão geral da arquitetura do *Swin Transformer*



Fonte: Liu *et al*. (2021).

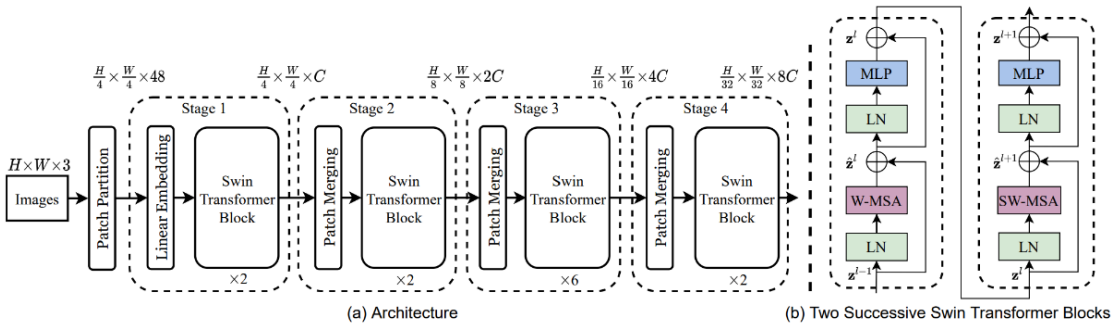
Neste trabalho utilizamos a versão *Swin Transformer Tiny* (ou SwinT). O modelo começa dividindo a imagem RGB de entrada em *patches* não sobrepostos, como no ViT. O vetor de características do *patch* é inicialmente obtido concatenando-se os valores RGB dos *pixels*.

Em resumo, o *Swin Transformer* consegue integrar as vantagens características das CNN em VC com a eficiente e poderosa arquitetura do ViT conforme destacado por Liu *et al*. (2021). Isso é possível porque a representação hierárquica pode alcançar a invariância de escala e a abordagem de janelas deslocadas pode transmitir informações de forma eficiente dentro da janela local, enquanto a *Merging Layer* é responsável por integrar as informações globais dos pixels.

A arquitetura *Swin Transformer* conseguiu superar os modelos CNN SOTA em várias tarefas de visão. Os principais pontos abordados em Liu, Shih e Zhong (2021) foram:

- arquitetura *Swin Transformer*: a arquitetura desse modelo integra *Transformers* e uma arquitetura de imagem hierárquica, superando limitações dos *Transformers* convencionais ao processar imagens de alta resolução. Ao invés de operar em *patches* fixos, o modelo usa "janelas deslocadas" para capturar contexto em diversas escalas. O método de processamento é feito em estágios através de blocos menores de imagem, o "*Swin Block*", permitindo a captura de informações contextuais em diferentes abstrações. Além disso, o *Swin Transformer* usa mecanismos de deslocamento e camadas de *pooling* para melhorar a eficiência e captura de longo alcance. A Figura 10 demonstra a arquitetura da rede *Swin Transformer* V1.

Figura 10 – Arquitetura *Swin Transformer*



Fonte: Liu, Shih e Zhong (2021).

- Janelas Deslocadas (*Shifted Windows*): o *Swin Transformer* aplica a autoatenção a pequenas janelas de pixels que são deslocadas pela imagem. Isso permite que o *Swin Transformer* processe imagens de maneira mais eficiente.
- Hierarquia de Camadas (*Layer Hierarchy*): o *Swin Transformer* aplica uma série de Transformers em diferentes resoluções, semelhante à forma como as CNN aplicam camadas convolucionais em diferentes resoluções. Isso ajuda a capturar características de alto e baixo nível.
- Fatoração da Autoatenção (*Self-Attention Factorization* – SAF): a técnica foi introduzida pelo *Swin Transformer* para diminuir a complexidade

computacional. Em vez de aplicar a autoatenção de maneira convencional, que tem uma complexidade quadrática em relação ao comprimento da sequência de entrada, o *Swin Transformer* a divide em duas etapas: atenção dentro de uma janela específica e atenção entre diferentes janelas. Essa abordagem é fundamental para tornar a atenção em *Transformers* mais viável computacionalmente para tarefas em grande escala, pois a técnica SAF consegue reduzir a complexidade para níveis lineares ou próximo disso.

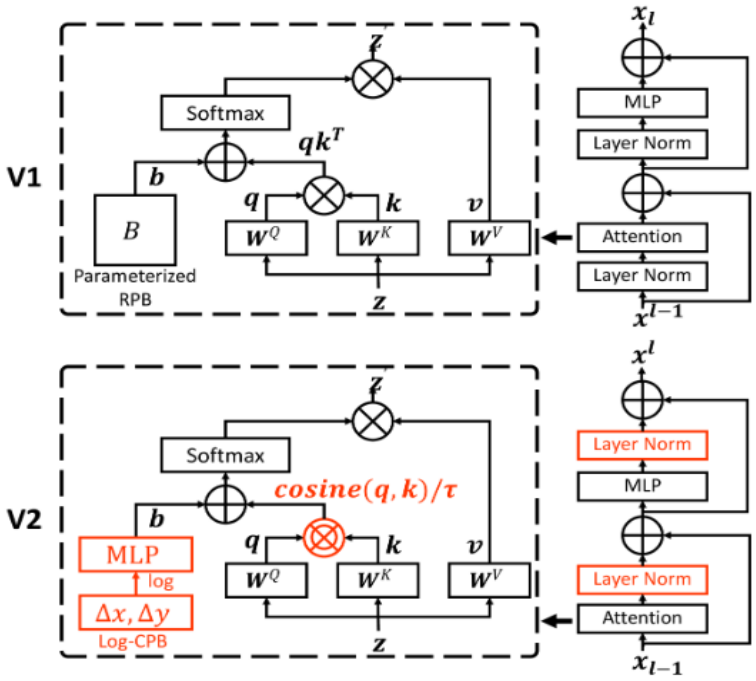
- adaptação para Tarefas de Visão: o *Swin Transformer* é projetado para ser facilmente adaptável para uma variedade de tarefas de visão, incluindo classificação de imagens, detecção de objetos e segmentação semântica.

2.2.2.5 Modelo *Swin Transformer* V2

Ainda explorando os avanços do modelo *Swin Transformer*, Liu, Shih e Zhong (2022) propuseram o modelo *Swin Transformer* V2. As principais características do *Swin Transformer* V2 são:

- estrutura de escala de imagem: o *Swin Transformer* V2 foi projetado para lidar com tarefas de VC em grande escala, como classificação de imagens e detecção de objetos em conjuntos de dados de alta resolução. Ele utiliza uma estrutura de escala de imagem que permite que a rede processe eficientemente informações em várias escalas, capturando detalhes finos e contextos globais.
- bloco *Swin*: o *Swin Transformer* V2 introduz um novo bloco chamado *Swin Block*, que substitui o bloco básico de atenção do *Transformer*. O bloco *Swin* utiliza uma estrutura hierárquica de divisão espacial para capturar informações contextuais de maneira mais eficiente. Ele contém uma camada de tokenização, uma camada de janela deslizante e um mecanismo de atenção deslocada. A Figura 11 demonstra a arquitetura da rede *Swin Transformer* V2 comparativamente à arquitetura do *Swin Transformer* V1.

Figura 11 – Arquitetura *Swin Transformer* V2



Fonte: Liu *et al.* (2022).

- ativação de escala variável: para melhorar a eficiência computacional e a escalabilidade, o *Swin Transformer* V2 utiliza ativações de escala variável em vez de ativações convolucionais tradicionais. Isso permite que a rede reduza o custo computacional enquanto mantém um desempenho promissor em tarefas de VC.
- estratégias de treinamento: o *Swin Transformer* V2 adota várias estratégias de treinamento para melhorar o desempenho do modelo. Isso inclui inicialização por camadas, um regime de treinamento multiescala e um novo esquema de programação de aprendizado.

Essas são apenas algumas das principais características do *Swin Transformer* V2. O modelo foi projetado para fornecer melhor desempenho em tarefas de VC em grande escala, aproveitando a estrutura hierárquica de divisão espacial do bloco *Swin* e as estratégias de treinamento aprimoradas.

Durante o estudo, Liu *et al.* (2022) identificaram três desafios principais na implementação do modelo — instabilidade de treinamento, diferenças de resolução entre pré-treinamento e ajuste fino, e a necessidade de grandes quantidades de dados rotulados.

Para superar a instabilidade no treinamento, os autores propuseram um método de "residual pós-norma" combinado com "atenção de cosseno". A nor-

malização residual é uma técnica comumente usada para estabilizar o treinamento de modelos de aprendizado profundo, enquanto a atenção de cosseno é uma variante da atenção baseada em produto escalar que utiliza a semelhança do cosseno em vez do produto escalar. Na sequência a questão das lacunas de resolução entre pré-treinamento e ajuste fino é abordada por meio de um "método de viés de posição contínua com espaçamento logarítmico". Isso sugere que o modelo é capaz de lidar com imagens de várias resoluções, um recurso valioso, dado que muitas tarefas de VC envolvem imagens de diferentes resoluções.

Para resolver a necessidade de grandes quantidades de dados rotulados, Liu *et al.* (2022) propõem o uso de um método de pré-treinamento auto supervisionado chamado *SimMIM*. A auto supervisão é uma técnica de aprendizado profundo em que os rótulos são gerados automaticamente a partir dos dados de entrada, permitindo que o modelo aprenda de grandes quantidades de dados não rotulados. Os autores afirmam ter treinado com sucesso um modelo *Swin Transformer* V2 de três bilhões de parâmetros. Dessa forma esse é o maior modelo de VC já desenvolvido e seria capaz de treinar com imagens de alta resolução (até 1.536 × 1.536). Esse modelo estabeleceu novos recordes de desempenho em várias tarefas de visão, incluindo classificação de imagem *ImageNet*-V2, detecção de objeto COCO, segmentação semântica *ADE20K* e classificação de ação de vídeo *Kinetics-400*,

Sobre o consumo de recursos, Liu *et al.* (2022) defendem que o treinamento desse modelo foi mais eficiente do que os modelos de VC de bilhões de parâmetros do *Google*, consumindo quarenta vezes menos dados rotulados e quarenta vezes menos tempo de treinamento. Esse é um resultado significativo, pois o treinamento de modelos de aprendizado profundo é frequentemente um processo intensivo em termos de recursos. Em resumo, este estudo representa um dos mais recentes avanços na área de VC de grande escala, introduzindo técnicas para superar desafios comuns e estabelecendo novos *benchmarks* de desempenho.

2.2.3 Aplicação dos Modelos de VC no Campo de Classificação de Imagens Radiográficas

A pandemia de COVID-19 teve sua origem em Wuhan, província de Hubei na China, em dezembro de 2019, e resultou em um esforço global para desenvolver métodos eficazes de detecção e diagnóstico da doença. As imagens de CXR surgiram como uma ferramenta valiosa neste processo, com muitos pesquisadores explorando técnicas de aprendizado profundo, como CNN e *Transformers*, para melhorar a velocidade e a acurácia da detecção.

Cheng *et al*. (2021), em artigo publicado pela revista *RadioGraphics*, vinculada à *Radiological Society of North America* (RSNA), destacam que os modelos de DL têm chamado a atenção da comunidade médica da área de radiologia de modo que já é considerado como adequada a utilização dessas ferramentas para apoio ao diagnóstico médico. Os autores, defendem que os modelos de VC provavelmente se tornarão ferramentas auxiliares para os médicos radiologistas. Dessa forma, a compreensão dos principais conceitos sobre os modelos de DL é primordial para que os radiologistas se mantenham informados sobre os avanços na área e haja maior facilidade na adoção clínica dessas técnicas.

Em sua revisão, Van Ginneken (2017) aborda a evolução do diagnóstico assistido por computador (da sigla em inglês, CAD) no diagnóstico a partir de imagem pulmonar, destacando a transição do processamento baseado em regras para o aprendizado de máquina e, mais recentemente, para o aprendizado profundo. As CNN são apresentadas como eficazes extratoras de recursos e classificadoras. Segundo os autores, elas são úteis para várias aplicações de CAD, incluindo a detecção de nódulos e a remoção de ruídos em imagens. O artigo também aponta o potencial das CNN para a análise de texto, sugerindo que a combinação de relatórios de texto e análise de imagem pode ser uma direção promissora para a pesquisa futura. Eles concluem que o DL tem um papel crucial a desempenhar no futuro da análise de imagens médicas, dada a sua capacidade superior para detectar padrões complexos e sutis nas imagens.

Esteva *et al*. (2019) proporcionaram um guia sobre o emprego do aprendizado profundo na área da saúde, salientando a importância dessas aplicações de para o diagnóstico e gestão de doenças.

Em suma, enquanto as CNN têm sido a espinha dorsal do progresso na VC, os *Transformers* estão se mostrando cada vez mais promissores e podem representar o futuro da VC.

Dessa forma, a aplicação desses modelos para apoiar os médicos em tarefas como classificação de casos de COVID-19 pode trazer grande auxílio para os serviços de saúde, como destacado no Subcapítulo 1.3.

2.3 TRABALHOS RELACIONADOS

2.3.1 Trabalhos Relacionados que Utilizaram Redes Neurais Convolucionais (CNN)

Na análise das bases científicas *Scopus* e *Web of Science*, identificaram-se artigos recentes sobre a classificação de imagens médicas para o diagnóstico da COVID-19. Embora muitos desses estudos tenham empregado modelos CNN, notou-se também um aumento na exploração dos modelos ViT e *Swin Transformer*. Estritamente sobre CNN selecionamos os trabalhos que seguem:

Castiglione *et al.* (2021) propuseram uma abordagem para a rápida detecção do COVID-19, utilizando um modelo otimizado de rede neural convolucional, chamado *ADECOCNN*. Tal modelo foi aplicado a imagens de TC do tórax. O *ADECOCNN* foi desenvolvido especificamente para diferenciar pacientes com COVID-19 dos não infectados tendo sido testado em comparação com outros modelos de CNN pré-treinados, — entre eles VGG19, *GoogleNet* e *ResNet* — que são reconhecidos na comunidade científica. Os resultados revelaram que o *ADECOCNN* supera esses modelos, obtendo uma acurácia de 0,9999, sensibilidade de 0,9996, precisão de 0,9992 e especificidade de 0,9997 na classificação de imagens de TC. Para garantir que o modelo possa ser aplicado em diferentes cenários e evitar o sobreajuste, foi empregada uma validação cruzada em cinco etapas. Apesar dos resultados encorajadores, os autores enfatizam a necessidade de desenvolver modelos que possam diferenciar COVID-19 de doenças semelhantes, como pneumonia, e sugerem que outros fatores de risco para o início da doença devem ser considerados para uma abordagem mais completa do problema.

Awan *et al*. (2021) empregaram o *Apache Spark* para implementar um método de *Deep Transfer Learning* (DTL) usando as arquiteturas de CNN *InceptionV3*, *ResNet50* e *VGG19* em radiografias torácicas de pacientes com COVID-19. Ao avaliar os modelos em duas categorias (radiografias normais e com COVID-19), alcançaram uma acurácia de 100%. Porém, ao incorporar a classe pneumonia, a acurácia foi de 0,97 para *InceptionV3*, e 0,9855 para *ResNet50* e *VGG19*. O sistema de classificação foi baseado em *pipelines* de aprendizado profundo e regressão logística, utilizando dois *datasets* do *Kaggle* para desenvolvimento e teste.

Banerjee *et al.* (2022) propuseram um *ensemble* de aprendizado profundo denominado Rede de Conjunto *Fuzzy* COVID, ou *COFE-Net*, para a triagem de COVID-19 a partir de CXR e tomografias computadorizadas. O estudo utilizou a estratégia de aprendizado por transferência para CNN e propôs um *ensemble* para combinar essas estratégias. Para combinar as pontuações de decisão geradas por três CNN (*Inception V3*, *Inception ResNet V2* e *DenseNet 201*), os princípios da lógica *fuzzy* foram aplicados através da integral *fuzzy* de *Choquet*.

Banerjee *et al.* (2022) atingiram uma acurácia de 0,9639 para classificação de três classes e 0,9949 para classificação de duas classes no conjunto de dados COVIDx. Além disso, também teve um desempenho promissor no conjunto de dados de *Montgomery County X-Ray* para a detecção de tuberculose, com uma acurácia de 0,9643. Isso demonstra que o método pode se adaptar a outros campos de imagens médicas. As limitações do método proposto por Banerjee *et al.* (2022) incluem a determinação empírica de medidas *fuzzy*, que é a base do mecanismo de combinação *fuzzy*. Outra limitação é que os classificadores CNN utilizados extraem características globalmente de todas as imagens de entrada, enquanto os elementos distintivos podem estar concentrados em uma parte específica da imagem. No futuro, os pesquisadores gostariam de usar um mecanismo de atenção para melhorar o foco nas regiões pulmonares afetadas na imagem para que recursos mais precisos possam ser extraídos.

Hamza *et al*. (2022) desenvolveram um *framework* com CNN otimizada pela técnica Bayesiana e IA explicável para classificar COVID-19 em imagens de CXR. Após realçar o contraste da imagem, eles adaptaram os modelos *EfficientNet-B0* e *MobileNet-V2*, treinando-os com otimização bayesiana. As características extraídas dos modelos foram combinadas por uma fusão serial e clas-

sificadas com algoritmos de aprendizado de máquina. A visualização foi aprimorada com o *Grad-CAM*, evidenciando áreas infectadas. Utilizando três datasets de COVID-19, alcançaram acurácias de 0,988, 0,979 e 0,994, validando a eficácia da abordagem. A principal limitação foi o aumento do tempo computacional após a fusão.

Kathamuthu *et al*. (2023) propuseram o desenvolvimento de alguns modelos CNN, com aprimoramento de aprendizado por transferência, para a detecção de COVID-19 em imagens de TC. Os autores testaram vários modelos de CNN com aprendizado por transferência, incluindo *VGG16*, *VGG19*, *Densenet121*, *InceptionV3*, *Xception* e *Resnet50*. Cada modelo foi avaliado utilizando uma matriz de confusão e várias medidas de desempenho, como acurácia, sensibilidade, precisão, F1-score, perda e ROC. Entre todos os modelos testados, o *VGG16* se destacou como o mais eficiente, alcançando uma acurácia de 0,98, além disso os autores destacam que outras vantagens do *VGG16* são que o modelo possui um menor número de parâmetros e menor tempo de treinamento.

Nishio *et al*. (2022) desenvolveram e validaram um modelo de aprendizado profundo para classificar imagens de CXR em pneumonia por COVID-19, pneumonia não-COVID-19 e indivíduos saudáveis. Utilizando o *EfficientNet* e mais de 20.000 imagens de dois conjuntos públicos e um privado, o modelo alcançou uma acurácia de 0,8667, superando a performance dos radiologistas, cujas acurácias variaram de 0,5667 a 0,7733. O desempenho do modelo foi particularmente notável na classificação de pneumonia por COVID-19, com um AUC de 0,9752 contra 0,8740 dos radiologistas. Apesar do sucesso, os autores identificaram limitações, como a necessidade de validação externa e a exclusão de outras doenças pulmonares. A ferramenta *Grad-CAM*, que poderia auxiliar na interpretação, não foi validada neste trabalho.

Srivastava *et al*. (2022) propuseram a utilização de IA para detectar COVID-19 através de imagens de CXR, usando um conjunto de dados curado com 1.281 imagens de COVID-19, 3.270 normais e 1.656 de pneumonia viral. Eles compararam diversos modelos pré-treinados, destacando-se o *InceptionV3* e *EfficientNet B0&B1* com acurácia de 0,9978 na classificação binária e o *ResNetV2* com acurácia de 0,9790 na classificação com três classes. Adicionalmente, criaram o *CoviXNet*, um novo modelo de CNN com 15 camadas, que obteve

acurácias de 0,9947 e 0,9661, respectivamente, para classificações binárias e tríplices. O *CoviXNet* combinou alta precisão com eficiência computacional.

Abiyev e Ismail (2021) apresentaram um modelo baseado em CNN para o diagnóstico de doenças de pneumonia viral COVID-19 e não-COVID-19. O modelo foi treinado em dois conjuntos de dados diferentes, sendo um para classificação binária (pneumonia/normal) e outro para classificação em três classes (COVID-19/pneumonia/normal) usando transferência de aprendizado. Os resultados obtidos foram promissores, com acurácia, sensibilidade, precisão e F1-score de 0,983, 0,979, 0,983 e 0,98, respectivamente, nos dados de teste. O modelo proposto demonstrou eficiência no diagnóstico de casos de COVID-19 e pneumonia, auxiliando os radiologistas no diagnóstico precoce dessas doenças. Segundo os autores, o artigo também explorou a visualização de mapas de ativação (*Grad-CAM*) para compreender o aprendizado do modelo e comparou os resultados com outros trabalhos, mostrando um desempenho superior.

Aggarwal *et al.* (2022) propuseram uma revisão abrangente sobre o diagnóstico da COVID-19 baseado na análise de imagens de CXR e TC usando técnicas de aprendizado profundo. O estudo destaca o uso de CNN como uma abordagem popular para a classificação de COVID-19. São resumidos vários estudos significativos que utilizaram DL para a classificação de COVID-19 a partir de imagens de CXR e CT, bem como as principais bases de dados utilizadas nesses estudos. São discutidos os desafios enfrentados pelas abordagens de DL atuais no diagnóstico de COVID-19, como a falta de dados e a necessidade de maior validação clínica. O estudo destaca a importância da interpretabilidade dos modelos de DL e sugere a colaboração entre médicos, radiologistas e engenheiros de IA para o desenvolvimento de soluções confiáveis e interpretáveis. Também são apontadas direções futuras de pesquisa, como a avaliação cruzada em diferentes conjuntos de dados e a disponibilização de códigos e dados para replicação e validação dos métodos propostos.

Ahamed *et al*. (2021) desenvolveram um modelo de detecção de casos de COVID-19 baseado em CNN, treinado com um conjunto de dados composto por TC e imagens de CXR. Os autores utilizaram uma arquitetura modificada do *ResNet50V2* como modelo de Aprendizado Profundo. O conjunto de dados foi coletado de várias fontes disponíveis publicamente e incluiu quatro classes: COVID-19 confirmado, casos normais e casos confirmados de pneumonia viral e

bacteriana. O conjunto de dados foi pré-processado com um filtro de nitidez antes de ser alimentado no modelo proposto. O modelo alcançou uma acurácia de 0,96452 para casos de quatro classes (COVID-19/Normal/Pneumonia bacteriana/Pneumonia viral), 0,97242 para casos de três classes (COVID-19/Normal/Pneumonia bacteriana) e 0,98954 para casos de duas classes (COVID-19/Pneumonia viral) usando imagens de CXR. O modelo obteve uma acurácia abrangente de 0,99012 para casos de três classes (COVID-19/Normal/Pneumonia adquirida na comunidade) e 0,9999 para casos de duas classes (Normal/COVID-19) usando imagens de TC do tórax. Segundo os autores, essa alta taxa de acurácia apresenta um recurso novo e potencialmente importante para permitir que os radiologistas identifiquem e diagnostiquem rapidamente os casos de COVID-19 com equipamentos básicos, mas amplamente disponíveis.

Asif *et al.* (2022) propuseram um modelo de CNN para o rápido e preciso diagnóstico de pacientes infectados pelo coronavírus da COVID-19, utilizando CXR e TC. Segundo os autores, para isso, foram avaliados diversos modelos pré-treinados de aprendizado profundo, como *InceptionV3*, *Xception*, *MobileNetV2*, *NasNet* e *DenseNet201*, para a classificação de imagens médicas. Os autores propuseram uma arquitetura de CNN superficial e leve, com baixa taxa de falsos negativos, para a classificação de imagens de CXR de pacientes. O conjunto de dados utilizado continha 2.541 imagens de CXR de dois bancos de dados públicos diferentes, com casos positivos confirmados de COVID-19 e casos saudáveis. Os resultados mostram que o modelo proposto alcançou uma acurácia máxima de 0,9968 e, mais importante, sensibilidade, especificidade e AUC de 0,9966, 0,9970 e 0,9998, respectivamente. O modelo proposto possuía menos parâmetros e baixa complexidade em comparação com outros modelos de aprendizado profundo. Segundo os autores, os resultados experimentais demonstram a superioridade do método proposto em relação aos métodos de referência do estado da arte.

Costa *et al*. (2022) apresentam uma revisão da literatura de artigos que se propuseram a apresentar modelos para detecção da COVID-19 a partir de imagens médicas torácicas, como CXR e TC. Os autores identificaram os cem artigos mais citados nesse campo de investigação (até a data de publicação), levando em consideração diversos aspectos, como o tipo de imagem explorada,

configurações de aprendizado, estratégias de segmentação, Inteligência Artificial Explicável (da sigla em inglês, XAI) e disponibilidade de conjuntos de dados e códigos. Segundo os autores a análise revelou uma predominância de métodos de aprendizado profundo em comparação com métodos rasos, sendo que muitos trabalhos utilizaram a transferência de aprendizado sem realizar ajuste fino. Houve também um desequilíbrio entre o número de trabalhos utilizando CXR e CT *scan*, com uma maior ênfase nas CXR. Os resultados alcançados por esses modelos mostraram avanços significativos na detecção da COVID-19, porém, a aplicação prática dessas estratégias ainda é limitada, com poucos artigos disponibilizando sistemas online e a falta de aplicação em cenários reais. Segundo os autores, a revisão contribui para identificar as principais abordagens adotadas e as lacunas a serem exploradas em pesquisas futuras para o desenvolvimento de soluções confiáveis e eficientes no diagnóstico da COVID-19 a partir de imagens médicas torácicas.

Kibriya e Amin (2023) propuseram um sistema automático de identificação e classificação de COVID-19 baseado em redes residuais (*ResNet*). Os autores realizaram diversas experimentações visando a detecção e classificação de COVID-19, utilizando as arquiteturas *ResNet18* e *ResNet50* para extrair características profundas de imagens de CXR. O classificador SVM foi empregado para classificar os vetores de características extraído. Segundo os autores, os resultados experimentais mostraram que o sistema proposto consegue detectar eficientemente a COVID-19 a partir de imagens de CXR, alcançando uma melhor acurácia de 0,973 utilizando *ResNet50*. Os autores destacaram que o conjunto de dados utilizado era desbalanceado, o que pode afetar a precisão das predições.

Lanjewar, Shaikh e Parab (2022) propuseram um sistema em tempo real baseado em CNN para a previsão da doença COVID-19 a partir de imagens de CXR. Segundo os autores, o modelo CNN implementado apresentou resultados exemplares, com acurácia de 0,9994 e acurácia de validação de 0,9881. Foram comparados diferentes modelos de CNN e DCNN (*ResNet50*, *VGG19*, *InceptionV3* e *Xception*). O modelo CNN proposto se mostrou mais eficiente e foi implantado na nuvem como uma plataforma de serviço (PaaS), permitindo a detecção rápida e eficiente da COVID-19. Conforme defendido pelos autores, a implementação do modelo em nuvem proporciona acessibilidade em dispositivos

móveis e reduz a carga dos médicos, acelerando os procedimentos de teste para pacientes com COVID-19. Os autores defendem que as limitações do estudo incluem a previsão apenas de pacientes com COVID-19 e pulmões saudáveis, que podem ser superadas pela integração de classificadores de aprendizado de máquina (como SVM, *Random Forest*, KNN) com as características extraídas pela CNN.

Yuan *et al.* (2023) propuseram um novo modelo de *CapsuleNet* mais leve chamado *DPDH-CapNet* para auxiliar na detecção da COVID-19 em imagens de CXR. O modelo utiliza convoluções de profundidade (D), convoluções pontuais (P) e convoluções dilatadas (D) para extrair características das imagens, capturando efetivamente as dependências locais e globais dos recursos patológicos da COVID-19. Segundo os autores, o modelo utiliza cápsulas vetoriais homogêneas (H) na camada de classificação, evitando a necessidade de cálculos complexos e caros de roteamento de cápsulas. Os resultados experimentais demonstram que o modelo *DPDH-CapNet* obteve uma acurácia de 0,9799, precisão de 0,9805, sensibilidade de 0,9802 e F1-score de 0,9803. Os autores ainda defendem que o modelo é mais leve em comparação com redes de cápsulas existentes, apresentando uma redução de nove vezes nos parâmetros. Além disso, defendem que o modelo não requer pré-treinamento e muitas amostras de treinamento. Os autores argumentam que o *DPDH-CapNet* contribui para uma compreensão mais profunda dos aspectos críticos dos casos de COVID-19 e mostra promessa como uma ferramenta para radiologistas e profissionais de saúde na detecção precisa da doença. Por fim, reconhecem, no entanto, ser necessário realizar pesquisas e testes clínicos adicionais para validar sua eficácia do modelo em um ambiente clínico real.

A Tabela 1 traz o resumo das métricas obtidas nos trabalhos selecionados na revisão da literatura sobre modelos CNN aplicados a imagens médicas para detectar COVID-19. Ressalta-se que os resultados encontrados, salvo algumas exceções, não são diretamente comparáveis entre si. Isso ocorre principalmente porque os datasets empregados são diferentes e porque não são fornecidas todas as métricas que comumente são empregadas para comparar dois modelos distintos.

Tabela 1 – Métricas dos modelos *ResNet50 (CNN)* relatadas na literatura

Referência	AUC	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score
Castiglione *et al*. (2021)	0,98	0,9999	0,9992	0,9996	0,9997	-
Awan *et al*. (2021)	-	1	-	-	-	-
	-	0,9700	-	-	-	-
	-	0,9855	-	-	-	-
Banerjee *et al*. (2022)	-	0,9639	-	-	-	-
	-	0,9949	-	-	-	-
Hamza *et al*. (2022)	-	0,9880	-	-	-	-
	-	0,9790	-	-	-	-
	-	0,9940	-	-	-	-
Kathamuthu *et al*. (2023)	-	0,9800	-	-	-	-
Nishio *et al*. (2022)	-	0,8667	-	-	-	-
	-	0,9752	-	-	-	-
	-	0,9492	-	-	-	-
	-	0,9912	-	-	-	-
Srivastava *et al*. (2022)	-	0,9978	-	-	-	-
	-	0,9790	-	-	-	-
	-	0,9947	-	-	-	-
	-	0,9661	-	-	-	-
Abiyev e Ismail (2021)	-	0,9830	0,9830	0,9790	-	0,9800
Ahamed *et al*. (2021)	-	0,9645	-	-	-	-
	-	0,9724	-	-	-	-
	-	0,9895	-	-	-	-
	-	0,9901	-	-	-	-
Asif *et al*. (2022)	0,9998	0,9968	-	0,9966	0,9970	-
Kibriya e Amin (2023)	-	0,9730	-	-	-	-
Lanjewar *et al*. (2022)	-	0,9881	-	-	-	-
Yuan *et al*. (2023)	-	0,9799	0,9805	0,9802	-	0,9803

Fonte: o autor (2023).

2.3.2 Trabalhos Relacionados que Utilizaram *Vision Transformer (*ViT)

Apesar dos avanços proporcionados pelas CNN, nos últimos anos elas têm sido desafiadas pelos modelos ViT originalmente criados para tarefas de NLP (Dosovitskiy *et al*., 2020), mas que têm se revelado promissores para aplicações no domínio de imagens.

Na revisão da literatura encontramos alguns trabalhos importantes reportando os resultados obtidos no emprego do ViT.

Shome *et al*. (2021) propuseram a utilização de um modelo ViT para a detecção de COVID-19 a partir de imagens de CXR. Foi utilizado um conjunto de dados agregado contendo trinta mil imagens, a maior coleção publicamente disponível neste domínio. O modelo demonstrou uma precisão de 0,98 e uma pon-

tuação AUC de 0,99 na classificação binária, e 0,92 de precisão e 0,98 de pontuação AUC na classificação multiclasse. O modelo superou outras arquiteturas conhecidas, como *EfficientNetB0*, *InceptionV3*, *Resnet50*, *MobileNetV3*, *Xception* e *DenseNet-121*, em todas as métricas. Além disso, os autores implementaram uma visualização baseada em *Grad-CAM* para tornar o modelo interpretável por radiologistas. Este trabalho ressalta o potencial dos CXR como uma ferramenta de diagnóstico valiosa e de baixo custo para a COVID-19, especialmente em situações em que testes rápidos são escassos.

Cao *et al.* (2022) propuseram uma estrutura de fusão de *CNN-Transformer* para a classificação automática de pneumonia em CXR, com ênfase na detecção da COVID-19. A estrutura consiste em duas partes: processamento de dados e classificação de imagens. O processamento de dados visa padronizar o formato de armazenamento dos dados de diferentes instituições médicas. Na etapa de classificação, é utilizada uma rede *multibranch* com um módulo de convolução personalizado e um módulo de *Transformer*. As sub-redes de extração de características extraem informações superficiais da imagem, enquanto as sub-redes de foco de características combinam características locais e globais. A rede proposta foi implementada e avaliada em conjuntos de dados de referência, alcançando uma acurácia de 97,09%, precisão de 97,16%, sensibilidade de 96,93% e F1-score de 97,04%. Comparada a outros métodos propostos, a rede obteve resultados superiores em termos de acurácia, precisão e F1-score na detecção da COVID-19. Os autores concluem que essa fusão de *CNN-Transformer* se mostrou promissora no diagnóstico automático de pneumonia, com destaque para a detecção da COVID-19, e há potencial para aprimoramentos futuros e aplicações clínicas mais amplas.

Chetoui e Akhloufi (2022) propuseram o uso de ViT para a detecção de COVID-19 em imagens de CXR. Vários modelos ViT foram ajustados para a tarefa de classificação multiclasse, distinguindo casos de COVID-19, Pneumonia e Normal. O conjunto de dados utilizado continha 7598 imagens de CXR com COVID-19, 8552 imagens de CXR de pacientes saudáveis e 5674 imagens de CXR de casos de pneumonia. Os resultados obtidos demonstraram um alto desempenho, alcançando AUC de 0,99 para a classificação multiclasse (COVID-19 *vs*. Outra Pneumonia *vs*. Normal). A sensibilidade alcançada para a classe COVID-19 foi de 0,99. O estudo mostrou que o modelo proposto superou arquitetu-

ras de redes CNN comparáveis no estado da arte na detecção de COVID-19 em imagens de CXR. Os autores desenvolveram vários modelos baseados em ViT, incluindo *ViT-B16*, *ViT-B32* e *ViT-L32*, para a detecção de COVID-19 em imagens de CXR. O modelo com melhor desempenho foi o ViT-B32, que alcançou AUC de 0,991, bem como uma especificidade e sensibilidade de 0,96. Especificamente para a classe COVID-19, obteve-se uma sensibilidade, especificidade e acurácia de 0,99, indicando um desempenho superior na identificação correta dos casos positivos de COVID-19. Segundo os autores os resultados obtidos neste estudo destacam o potencial dos ViT em superar redes CNN profundas na detecção precisa de COVID-19 em imagens de CXR.

Park *et al*. (2022) propuseram uma abordagem chamada *Distillation for Self-Supervision and Self-Training Learning* (DISTL) inspirada no processo de aprendizado de radiologistas, com o objetivo de aprimorar o desempenho dos ViT por meio de auto supervisão e autoaprendizagem através da destilação de conhecimento. Os resultados obtidos na validação externa utilizando três hospitais para diagnóstico de tuberculose, pneumotórax e COVID-19 demonstraram um desempenho progressivamente melhor à medida que a quantidade de dados não rotulados aumentava, superando até mesmo o modelo totalmente supervisionado com a mesma quantidade de dados rotulados. A AUC foi utilizada como a métrica principal de avaliação, e os valores obtidos de 0,99 para a classificação multiclasse (COVID-19 *vs*. outra pneumonia *vs*. normal). Além disso, a sensibilidade, acurácia e especificidade obtidas foram de 0,99 para a classe COVID-19. Segundo os autores os resultados evidenciam que a abordagem proposta supera modelos de referência comparáveis na detecção de COVID-19 em imagens de CXR, utilizando arquiteturas de redes CNN profundas.

Park, Choi e Lee (2022) propuseram uma abordagem inovadora para desenvolver um algoritmo robusto de diagnóstico e quantificação da gravidade da COVID-19 usando CXR. O modelo *Multi-task ViT* alcançou um desempenho de ponta, com métricas elevadas, tanto no diagnóstico quanto na quantificação da gravidade da doença. A AUC foi de 0,998 para a classificação multiclasse. Segundo os autores, o modelo apresentou uma capacidade notável de generalização, demonstrando estabilidade em diferentes conjuntos de testes externos, independentemente das configurações e visões das CXR. O método proposto também se mostrou promissor como uma ferramenta de triagem, sendo capaz

de priorizar de forma confiável a população com baixo risco de infecção, poupando recursos médicos ao evitar testes moleculares desnecessários. O modelo obteve *Negative Predictive Value* (NPV) superior a 99%, dessa forma, os autores defendem que o modelo poderia poupar da necessidade de realizar teste RT-PCR até 80% da população. Os autores ainda defendem que a estimativa da gravidade da infecção fornecida pelo modelo oferece orientação para decisões de tratamento e avaliação da resposta.

Marefat *et al.* (2023) apresentaram um método baseado em ViT para detectar automaticamente a COVID-19 a partir de imagens de Raio X empregando *Transformers Convolutivos Compactos* (CCT). Segundo os autores, a principal vantagem do CCT sobre outros modelos baseados em Transformers é a menor necessidade de dados, o que é crucial dada a escassez de dados em muitos domínios médicos. O estudo demonstrou a eficácia do modelo proposto na detecção de COVID-19, alcançando uma precisão de 99,22.

Mehboob *et al*. (2022) apresentaram uma abordagem baseada em autoatenção do ViT para diagnosticar a COVID-19, que foi avaliada em comparação com as CNN e classificadores *Ensemble*. Foram utilizadas imagens de TC dos conjuntos de dados binários *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2) e multiclasse *Hybrid-learning for UnbiaSed predicTion* of COVID-19 (HUST-19); o novo método atingiu uma alta precisão de 0,997 e 0,98, respectivamente. Além disso, a avaliação *cross-corpus* usando o conjunto de dados COVID HUST-19 retornou uma precisão de 0,93, demonstrando a eficácia da abordagem. O estudo também descobriu que algoritmos baseados em ensemble, na abordagem proposta e em CNN, tiveram desempenho mais preciso no conjunto de dados HUST19 do que no conjunto de dados brasileiro (SARS-CoV-2). Segundo destacaram os autores a abordagem do modelo ViT proposto pôde prever a gravidade da COVID-19, potencialmente auxiliando os clínicos a tomarem decisões informadas em relação ao cuidado do paciente.

Murphy *et al.* (2022) apresentaram um estudo retrospectivo que comparou a eficácia dos *Transformers* Visuais Eficientes em Dados de Imagem (DeiT) e das CNN, nomeadamente *DenseNet121*, *DeiT-Ti* (*Tiny*), *ResNet152* e *EfficientNetB7*, no diagnóstico de doenças em radiografias de tórax e de extremidades. Usando o conjunto de dados do NIH CXR 14 e o MURA, observou-se que a área ponderada sob a curva ROC (wAUC) para *DeiT-B* foi ligeiramente inferior à do

DenseNet121 nas radiografias de tórax (0,78 *vs*. 0,79) e de extremidades (0,887 *vs*. 0,893). O modelo DeiT-B apresentou uma menor prevalência de tubos torácicos em falsos-positivos de pneumotórax [43.1% (324 de 5.088)] em comparação com a *DenseNet121* [47.9% (2.290 de 4.782)], sugerindo uma menor suscetibilidade à estratificação oculta. Segundo os autores, apesar de os ViT apresentarem melhorias sobre as CNN em conjuntos de dados gerais de imagens, os resultados sugerem que ainda não estão prontos para substituir as CNN no diagnóstico radiográfico de doenças.

Jiang *et al.* (2022) propuseram uma nova variante do ViT Piramidal, chamado MXT, para a classificação de múltiplos rótulos em imagens de CXR. Segundo os autores, através da autoatenção o MXT consegue capturar informações visuais de curto e longo alcance nas imagens de Raio X. A abordagem usa atenção de redução espacial de subamostragem para reduzir o consumo de recursos do *Transformer*. Os resultados experimentais em dois conjuntos de dados demonstram que o MXT é eficiente para a classificação de múltiplos rótulos em imagens de CXR, apresentando uma AUC média de 0,83 no conjunto de dados CXR 14 e 0,946 no conjunto de dados *Catheter*.

Mondal *et al.* (2022) propuseram o uso de ViT para a detecção de COVID-19 em imagens de CXR e TC. O método, chamado xViTCOS, segundo os autores, superou os métodos baseados em CNN obtendo precisão, sensibilidade e F1-score de 0,981, especificidade de 0,992 e NPV de 0,991. Os autores avaliam que os resultados quantitativos mostram que o xViTCOS supera as referências recentes em métricas de desempenho. Além disso, o estudo analisa os casos em que o método falha e sugere o uso do xViTCOS em conjunto com o teste RT-PCR para aprimorar a eficácia do diagnóstico.

Konwer e Prasanna (2022) discutem que a análise automatizada de imagens de tórax em casos de COVID-19 tem sido limitada devido ao tamanho reduzido dos conjuntos de dados disponíveis, levando a problemas de *overfitting* e baixa generalização. Dessa forma os autores propõem uma abordagem baseada em ViT auto supervisionado para prever desfechos clínicos em CXR. Os autores utilizaram tarefas de reconstrução e aprendizado contrastivo para pré-treinar o ViT em um grande conjunto de dados não rotulados. Em seguida, extraíram características relevantes para a tarefa de predição utilizando um conjunto de dados de interesse.

Wang *et al*. (2023) apresentam o *PneuNet*, um modelo baseado no VIT que utiliza atenção baseada em canais para alcançar um diagnóstico preciso de pneumonia. Segundo os autores, ao aplicar atenção *multi-head* em *patches* de canal, o *PneuNet* supera os modelos de aprendizado profundo anteriores, alcançando uma precisão de 0,9496 no problema de classificação em três categorias. Além disso, o modelo proposto, baseado no *ResNet18*, demonstra alta precisão (0,9513) e acurácia (0,9516) na detecção de casos de COVID-19 a partir de imagens de CXR. Segundo os autores o *PneuNet* apresentou desempenho promissor na classificação binária, distinguindo COVID-19 de não pneumonia com uma acurácia de treinamento de 0,9929 e precisão de 0,9879. O modelo apresentou, ainda, resultados promissores (0,8694 de acurácia) em uma classificação de quatro categorias, incluindo COVID-19, não-pneumonia, pneumonia bacteriana e pneumonia viral.

Chen *et al.* (2023) propuseram um novo modelo de rede de aprendizado profundo (*BoT-ViTNet*) para classificação automática, baseado no *ResNet50*, primeiramente, introduzindo o mecanismo de MSA nas últimas camadas do bloco *Bottleneck* das três primeiras etapas do *ResNet50*, a fim de aprimorar a capacidade de modelar informações globais. Em seguida, para aprimorar ainda mais o desempenho na coleta de características e a correlação entre elas, os autores utilizam os blocos TRT-ViT, compostos por *Transformer* e *Bottleneck*, na etapa final do *ResNet50*, o que melhorou o reconhecimento de regiões de lesões complexas nas imagens de CXR. Segundo os autores, por fim, as características extraídas são entregues à camada de média global para integração das informações espaciais globais de forma concatenada e utilizadas para a classificação. Os experimentos conduzidos no banco de dados *COVID-19 Radiography* mostraram que a precisão, sensibilidade, especificidade, F1-score e acurácia de classificação do modelo *BoT-ViTNet* foram, respectivamente, de 0,9891, 0,9780, 0,9876, 0,9913 e 0,9827, superando outros modelos de classificação.

A Tabela 2 traz o resumo das métricas obtidas nos trabalhos encontrados na revisão da literatura. Ressalta-se que os resultados encontrados, salvo algumas exceções, não são diretamente comparáveis entre si. Isso ocorre principalmente porque os *datasets* empregados são diferentes e porque não são for-

necidas todas as métricas que comumente são empregadas para comparar dois modelos distintos.

Tabela 2 – Métricas dos modelos *Vision Transformer (ViT)* relatadas na literatura

Estudo	AUC	Acurácia	Precisão	Sensibilidade	Especificidade	F1-score
Shome *et al*. (2021)	0,99	0,98	-	-	-	-
Cao *et al*. (2022)	-	0,9709	0,9716	0,9693	-	0,9704
Chetoui e Akhloufi (2022)	0,99	-	-	0,99	-	-
Park *et al*. (2022a)	0,99	-	-	0,99	-	-
Park *et al*. (2022b)	1	-	-	-	-	-
Murphy *et al*. (2022)	0,78	-	-	-	-	-
Jiang *et al*. (2022)	0,83	-	-	-	-	-
Mehboob *et al*. (2022)	-	-	0,997	-	-	-
Mondal *et al*. (2022)	-	-	0,981	-	0,992	-
Konwer e Prasanna (2022)	-	-	-	-	-	-
Wang *et al*. (2023)	-	-	0,9513	-	-	-
Chen *et al*. (2023)	-	0,9827	0,9891	0,978	0,9876	0,9913
Marefat *et al*. (2023)	-	0,9922	-	-	-	-

Fonte: o autor (2023).

2.3.3 Trabalhos Relacionados que Utilizaram *Swin Transformer*

A arquitetura *Swin Transformer* foi proposta por Liu *et al.* (2021 e tem apresentado desempenho de ponta para tarefas de VC. Apesar de ser derivado do ViT (Dosovitskiy *et al*., 2020), o *Swin Transformer* é mais eficiente e tem maior precisão (Liu *et al*., 2021). Devido a essas propriedades desejáveis, o *Swin Transformer* pode ser usado como o *backbone* em muitas arquiteturas de modelos de VC. O *Swin Transformer* se propõe a resolver os problemas inerentes ao modelo ViT original usando mapas de recursos hierárquicos e MSA sobre a janela deslocada (*Shifted Windows*) (Liu *et al*. 2021). Na revisão da literatura foram selecionados alguns artigos científicos sobre *Swin Transformer,* conforme será detalhado em seguida.

Dinh *et al.* (2022) demonstraram que o *Swin Transformer* obteve precisão de 0,99 em tarefa de classificação para a classe COVID-19. Nesse artigo, os autores exploram o uso de métodos de aprendizado profundo para classificar

três tipos de CXR — normais, pneumonia e COVID-19. Cinco modelos de aprendizado profundo — *DenseNet121*, *ResNet50*, *InceptionNet*, *Swin Transformer* e *Hybrid EfficientNet-DOLG* — foram utilizados para realizar os experimentos em um conjunto de dados personalizado. Os autores também realizam um experimento na avaliação da gravidade do COVID-19. Os resultados indicam que a combinação de CXR e aprendizado profundo pode oferecer uma abordagem confiável para apoiar os médicos no diagnóstico e avaliação da gravidade do COVID-19. No entanto, mais pesquisas são necessárias para melhorar a robustez desses modelos e garantir sua eficácia em uma ampla gama de cenários clínicos.

Peng *et al*. (2022) propuseram o modelo *DeepDSR* como um *ensemble* contendo os modelos *DenseNet*, *Swin Transformer* e *RegNet*. Nesse estudo foi proposta uma estrutura de DL, chamada *DeepDSR*, que combina três arquiteturas de aprendizado profundo notáveis: *DenseNet*, *Swin Transformer* e *RegNet*. O estudo começou com a integração de três conjuntos de dados de imagens de TC relacionados à COVID-19 em um único conjunto maior. Em seguida, os pesos de *DenseNet*, *Swin Transformer* e *RegNet* foram pré-treinados no conjunto de dados *ImageNet* com base no aprendizado do *Transformer*. Os modelos foram então treinados no conjunto de dados de imagem maior e integrado. Por fim, os resultados da classificação foram obtidos integrando os resultados dos três modelos acima e usando uma abordagem de *soft voting*. O modelo *DeepDSR* proposto foi comparado com três modelos de aprendizado profundo de última geração — *EfficientNetV2*, *ResNet* e *ViT* — e com os três modelos individuais (*DenseNet*, *Swin Transformer* e *RegNet*) para problemas de classificação binária e de três classes. Os resultados mostraram que o *DeepDSR* obteve as melhores métricas: precisão de 0,9833, sensibilidade de 0,9895, acurácia de 0,9894, *F1-score* de 0,9864, AUC de 0,9991 e AUPR de 0,9986. Estes resultados sugerem a eficácia do *DeepDSR* na identificação de COVID-19, indicando seu potencial para contribuir significativamente para o diagnóstico da COVID-19.

Tian *et al*. (2022) propuseram um modelo para a detecção rápida de pacientes com COVID-19 por meio de imagens de TC pulmonar usando tecnologia de aprendizado profundo. O modelo proposto, VitCNX, combina as arquiteturas ViT e *ConvNeXt* para identificação de imagens de TC de COVID-19. Os resulta-

dos obtidos foram comparados com os de outros modelos, como *EficientNetV2*, *DenseNet*, *ResNet*-50 e *Swin Transformer*, além dos modelos individuais utilizados no conjunto. Os experimentos realizados envolvem classificações binárias e de três classes. No caso da classificação binária, o VitCNX obteve os melhores resultados em termos de sensibilidade (0,9907), acurácia (0,9821), F1-score (0,9855), AUC (0,9985) e AUPR (0,9991), superando os outros modelos. Da mesma forma, no experimento de três classificações, o VitCNX demonstrou uma excelente capacidade de classificação de imagem, alcançando uma precisão de 0,9668, uma acurácia de 0,9696 e uma pontuação F1-score de 0,9631. Os autores defendem que esses resultados sugerem que o modelo VitCNX pode ser uma ferramenta promissora para o reconhecimento de pacientes com COVID-19 por meio de imagens de TC pulmonar. A combinação dos algoritmos ViT e ConvNeXt mostra-se eficaz na identificação precisa dos pacientes.

Ma e Lv (2022) aplicam o *Swin Transformer* para reconhecimento de pneumonia em imagens de CXR, sendo o modelo otimizado de acordo com as características dessas imagens. Os resultados experimentais com base no modelo proposto neste artigo são comparados com os resultados de um modelo construído com uma CNN tradicional como *backbone*, e a precisão do modelo é, segundo os autores, significativamente melhorada. Após os experimentos de comparação em dois conjuntos de dados diferentes, os resultados experimentais mostram que a precisão do modelo melhorou de 0,763 para 0,873 e de 0,928 para 0,972, respectivamente. Os autores aplicaram *Grad-CAM* para encontrar as regiões aproximadas correspondentes às lesões.

Pan *et al*. (2023) propuseram um *framework*, denominado *Denoising Diffusion Probabilistic Model* (MT-DDPM), para síntese de imagens médicas utilizando uma rede baseada em *Swin Transformer* e um modelo de difusão para abordar a limitação de conjuntos de dados limitados no treinamento de modelos de VC em imagens médicas. As avaliações visuais e quantitativas demonstraram que as imagens sintéticas geradas possuíam uma aparência visual realista, com alta qualidade e diversidade. Na tarefa de classificação de COVID-19, a utilização de uma combinação de dados sintéticos e reais resultou em precisão de 0,93, superando a precisão obtida utilizando apenas dados reais ou sintéticos. Dessa forma, foi evidenciado o potencial do *framework* para complementar

conjuntos de treinamento limitados e fornecer uma solução promissora para aplicações de pesquisa em imagens médicas.

Para facilitar a comparação das métricas obtidas nos estudos sobre *Swin Transformer* foi elaborada a Tabela 3. Ressalta-se que os resultados encontrados, salvo algumas exceções, não são diretamente comparáveis entre si. Isso ocorre principalmente porque os datasets empregados são diferentes e porque não são fornecidas todas as métricas que comumente são empregadas para comparar dois modelos distintos.

Tabela 3 – Métricas dos modelos *Swin Transformer* relatadas na literatura

Modelo/Estudo	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**Especificidade**	**AUPR**	**F1-score**
Swin Transformer (Dinh *et al*., 2022)	-	-	0,99	-	-	-	-
DeepDSR (Peng *et al*., 2022a)	0,9991	0,9894	0,9833	0,9895	-	0,9986	0,9864
VitCNX (Tian *et al*., 2022) (Binária)	0,9985	0,9821	-	0,9907	-	0,9991	0,9855
VitCNX (Tian *et al*., 2022) (três classes)	-	0,9696	0,9668	-	-	-	0,9631
Swin Transformer (Ma e Lv, 2022) (Conjunto 1)	-	0,873	-	-	-	-	-
Swin Transformer (Ma e Lv, 2022) (Conjunto 2)	-	0,972	-	-	-	-	-
MT-DDPM (Pan *et al*., 2023)	-	0,93	-	-	-	-	-

Fonte: o autor (2023).

Os estudos recentes na área de imagens médicas têm mostrado um contínuo aprimoramento, explorando diferentes modelos como CNN puro, *Transformer* e *Swin Transformer*, ou até mesmo a combinação deles em *ensembles*. Além disso, os pesquisadores estão cada vez mais interessados em melhorar as arquiteturas e alcançar métricas mais robustas para aplicação em imagens médicas. Esses avanços refletem o compromisso em impulsionar o progresso e a qualidade das soluções desenvolvidas para diagnósticos e análises médicas.

2.3.4 Consolidação dos Resultados Encontrados na Literatura Para as Três Arquiteturas

Inicialmente é importante analisar os resultados encontrados na revisão da literatura que foi realizada e encontram-se compilados na Tabela 4. Como foi

elucidado anteriormente, a revisão da literatura buscou avaliar artigos publicados em periódicos de ponta sobre a aplicação dos modelos CNN, ViT e SwinT para o problema de classificação de imagens médicas e auxiliar no diagnóstico da COVID-19.

Tabela 4 – Resultados encontrados na revisão de literatura

Modelo	Autor	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	Castiglione *et al*. (2021)	0,98	0,9999	0,9992	0,9996	-
	Awan *et al*. (2021)	-	0,97	-	-	-
	Banerjee *et al*. (2022)	-	0,9949	-	-	-
	Hamza *et al*. (2022)	-	0,994	-	-	-
	Kathamuthu *et al*. (2023)	-	0,98	-	-	-
	Nishio *et al*. (2022)	-	0,9912	-	-	-
	Srivastava *et al*. (2022)	-	0,9978	-	-	-
	Abiyev e Ismail (2021)	-	0,983	0,983	0,979	0,98
	Ahamed *et al*. (2021)	-	0,9901	-	-	-
	Asif *et al*. (2022)	0,9998	0,9968	-	0,9966	-
	Kibriya e Amin (2023)	-	0,973	-	-	-
	Lanjewar *et al*. (2022)	-	0,9881	-	-	-
	Yuan *et al*. (2023)	-	0,9799	0,9805	0,9802	0,9803
	Shome *et al*. (2021)	0,99	0,98	-	-	-
	Cao *et al*. (2022)	-	0,9709	0,9716	0,9693	0,9704
ViT	Chetoui e Akhloufi (2022)	0,99	-	-	0,99	-
	Park *et al*. (2022a)	0,99	-	-	0,99	-
	Park *et al*. (2022b)	0,998	-	-	-	-
	Murphy *et al*. (2022)	0,78	-	-	-	-
	Jiang *et al*. (2022)	0,83	-	-	-	-
	Mehboob *et al*. (2022)	-	-	0,997	-	-
	Mondal *et al*. (2022)	-	-	0,981	-	-
	Konwer e Prasanna (2022)	-	-	-	-	-
	Wang *et al*. (2023)	-	-	0,9513	-	-
	Chen *et al*. (2023)	-	0,9827	0,9891	0,978	0,9913
	Marefat *et al*. (2023)	-	0,9922	-	-	-
SwinT	Dinh *et al*., (2022)	-	-	0,99	-	-
	Peng *et al*., (2022)	0,9991	0,9894	0,9833	0,9895	0,9864
	Tian *et al*., (2022) (Binária)	0,9985	0,9821	-	0,9907	0,9855
	Tian *et al*., (2022) (Três classes)	-	0,9696	0,9668	-	0,9631
	Ma & Lv (2022) (Conjunto 1)	-	0,873	-	-	-
	Ma & (Lv 2022) (Conjunto 2)	-	0,972	-	-	-
	Pan *et al*., (2023)	-	0,93	-	-	-

Fonte: o autor (2023).

Para facilitar o entendimento sobre quais bases de dados que foram utilizadas pelos trabalhos, foram compilados os resultados organizados por modelo no Quadro 2, abaixo:

Quadro 2 – Bases de dados utilizadas pelos artigos da revisão de literatura

Modelo	Autor	Conjunto de Dados
CNN	Castiglione *et al*. (2021)	*SARS-COV-2 CT-Scan*
	Awan *et al*. (2021)	*Coronavirus chest x-ray images and Chest X-Ray images (Pneumonia)*
	Banerjee *et al*. (2022)	*SARS-COV-2 CT Scan Dataset and Montgomery Dataset (CXRs)*
	Hamza *et al*. (2022)	*COVID-GAN, COVID-Net small chest x-ray and CXR*
	Kathamuthu *et al*. (2023)	*Several*
	Nishio *et al*. (2022)	*COVID-BIMCV and COVID-PRIVATE*
	Srivastava *et al*. (2022)	*Chest X-rays and CT Scan Dataset*
	Abiyev e Ismail (2021)	*CXR images and COVID-19 Radiography database*
	Ahamed *et al*. (2021)	*COVID-19 Radiography Database*
	Asif *et al*. (2022)	*COVID-19 radiography database*
	Kibriya e Amin (2023)	*Chest X-ray*
	Lanjewar *et al*. (2022)	*COVID-19 Radiography Database and Chest X-Ray Images*
	Yuan *et al*. (2023)	*Covid chest x-ray and Mixed Dataset*
ViT	Shome *et al*. (2021)	*Mixed Dataset*
	Cao *et al*. (2022)	*Guangzhou Women and Children Medical Centre dataset, MIDRC-RICORD and COVIDx CXR dataset*
	Chetoui e Akhloufi (2022)	*ChestX-ray8 dataset, NIH dataset and RSNA dataset*
	Park *et al*. (2022a)	*CheXpert, BIMCV, PADChest and Montgomery and Shenzen dataset*
	Park *et al*. (2022b)	*CheXpert, BIMCV, Brixia, NIH, AMC, CNUH, YNU, KNUH datasets*
	Murphy *et al*. (2022)	*Chest X-ray 14 dataset and MURA dataset*
	Jiang *et al*. (2022)	*MXT dataset and Chest X-ray14 dataset*
	Mehboob *et al*. (2022)	*HUST-19 dataset*
	Mondal *et al*. (2022)	*COVIDx CT-2A dataset and Chest X-RAY Dataset*
	Konwer e Prasanna (2022)	*COVID-19 Radiography Database, SIIM-FISABIO-RSNA COVID-19 Detection*
	Wang *et al*. (2023)	*Mixed 7 datasets*
	Chen *et al*. (2023)	*COVID-19 Radiography database*
	Marefat *et al*. (2023)	*COVIDx CXR-3*
SwinT	Dinh *et al*. (2022)	*COVID CXR, Chest X-ray, RICORD dataset and RALO dataset*
	Peng *et al*. (2022)	*COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT*
	Tian *et al*. (2022) (Binária)	*COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT*
	Tian *et al*. (2022) (Três classes)	*COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT*
	Ma & Lv (2022) (Conjunto 1)	*NIH dataset and Chest X-Ray dataset*
	Ma & Lv (2022) (Conjunto 2)	*NIH dataset and Chest X-Ray dataset*
	Pan *et al*. (2023)	*NIH Chest x-rays dataset, MRI dataset and CT dataset*

Fonte: o autor (2023).

Sobre os resultados compilados dos 33 artigos, pode-se verificar que os modelos CNN, ViT e SwinT, aplicados na classificação de COVID-19 e NORMAL, apresentaram resultados promissores em termos de valores absolutos das métricas obtidas. No entanto, a comparação direta entre os modelos e mesmo os resultados de um mesmo modelo devem ser realizadas com muito cuidado, uma vez que as particularidades dos *datasets* utilizados em cada trabalho, bem como as diferenças de abordagem e arquiteturas das redes utilizadas é uma limitação que impede que uma comparação de uma métrica frente a outra possa ser determinante para a decisão se um modelo é melhor ou pior que outro.

Pode-se observar que os modelos CNN tiveram resultados promissores, com altos níveis de acurácia e precisão, apesar de algumas discrepâncias nos resultados de diferentes estudos. A variação de desempenho entre os modelos CNN pode ser atribuída a diferenças nas arquiteturas de modelo, técnicas de pré-processamento de dados e conjuntos de dados utilizados.

Os modelos ViT e SwinT também foram explorados, com alguns estudos reportando resultados favoráveis, embora nem todos tenham fornecido informações completas sobre todas as métricas avaliadas. Não é possível comparar diretamente os modelos CNN e ViT devido à inconsistência das métricas disponíveis em cada estudo.

Na literatura levantada, as métricas documentadas para modelos de CNN geralmente enfatizam a acurácia, com pouca relevância dada à AUC e Sensibilidade. Essa tendência também é observada em pesquisas sobre modelos ViT, embora haja uma ligeira ênfase maior na AUC e Sensibilidade em alguns estudos. O mesmo padrão se repete em pesquisas sobre o modelo SwinT, com mais trabalhos abordando de forma um pouco mais acentuada a AUC e Sensibilidade.

Em resumo, todos os três modelos apresentam potencial para a classificação binária, mas há uma necessidade de informações padronizadas e mais detalhadas sobre as métricas para uma comparação adequada. É importante considerar a variação dos resultados entre os estudos e realizar pesquisas adicionais para confirmar e comparar a eficácia dos modelos em diferentes conjuntos de dados.

Tendo esse apontamento em mente, os experimentos do presente trabalho se propuseram em explorar qual seria o modelo mais adequado para a clas-

sificação de imagens médicas. Para tanto, no estudo, os mesmos *datasets* e condições de execução foram adotados para que uma avaliação mais justa, quanto possível, pudesse ser realizada.

2.4 A IMPORTÂNCIA DOS DATASETS PARA O *DEEP LEARNING* APLICADO ÀS IMAGENS MÉDICAS

Os últimos anos testemunharam um aumento substancial no uso de técnicas de DL para a análise de imagens médicas. Esse é um campo que demanda grandes conjuntos de dados para que o treinamento do modelo possa ser realizado com sucesso. Isso cresce de importância porque envolve a aplicação de técnicas avançadas de aprendizado de máquina para a interpretação e análise de dados de imagens médicas, com aplicações potenciais em diagnóstico, prognóstico e planejamento de tratamento.

Litjens *et al*. (2017) realizaram uma revisão abrangente do uso de DL na análise de imagens médicas. Eles destacaram a capacidade única de DL para capturar características complexas em dados de alta dimensão, tornando-a particularmente adequada para análise de imagens médicas. No entanto, eles também notaram que, apesar do sucesso inicial, ainda há muitos desafios a serem enfrentados no campo, incluindo a necessidade de conjuntos de dados de treinamento maiores e mais variados e a interpretação de modelos de aprendizado profundo.

Catala *et al*. (2021) exploraram a questão do viés em conjuntos de dados públicos de imagens de Raio X de pacientes com pneumonia e COVID-19. Este estudo destaca a importância de considerar o viés nos dados de treinamento ao desenvolver e avaliar modelos de aprendizado de máquina e DL na análise de imagens médicas.

Complementarmente, Arias-Garzón *et al.* (2023) discutem os desafios e possíveis vieses no uso de bases de dados de imagens de CXR para a detecção e estudo da COVID-19 usando algoritmos de IA. O estudo avaliou 19 *datasets* para emitir suas conclusões. Em resumo, principais dos pontos defendidos pelos autores são:

1. novidade do vírus e vieses associados: o surgimento recente da COVID-19 levou à rápida liberação de bases de dados para sua detecção e

estudo. No entanto, essas bases de dados podem conter vieses relacionados às informações do paciente, condições de captura, desequilíbrio e mistura de bases de dados.

2. análise de dados e metadados: é essencial realizar uma análise aprofundada dos dados e seus metadados para identificar possíveis vieses. Ferramentas éticas, como o *Aequitas*, podem ser usadas para verificar a presença de viés de idade, gênero ou raça, por exemplo.

3. construção de novas bases de dados: ao criar bases de dados, é recomendado que sejam o mais homogêneas possível para evitar a introdução de novos vieses. A mistura de bases de dados pode não ser recomendada, pois pode introduzir viés adicionais.

4. vieses ao misturar conjuntos de dados: misturar conjuntos de dados de COVID-19 com outros pode resultar em um modelo de IA que diferencia com base no conjunto de dados utilizado em vez do conteúdo da imagem. Isso pode resultar em uma distribuição desigual de características como idade, gênero e localização.

5. foco na área de interesse da doença: foi proposto que se realize a segmentação do pulmão em bases de dados existentes para ajudar a IA a identificar características específicas da doença e mitigar possíveis vieses.

6. recomendações para futuros estudos e criação de bases de dados: entre as recomendações estão a criação de dados homogêneos, a validação dos resultados com um radiologista especialista, a criação de metadados detalhados para as imagens e a evitar o uso de imagens de Unidades de Terapia Intensiva (UTI), se possível.

7. uso de ferramentas ou *frameworks* de IA éticos: é recomendado o uso de ferramentas ou *frameworks* de IA éticos para identificar possíveis vieses no modelo.

Por fim, Arias-Garzón *et al.* (2023) defendem que a identificação de COVID-19 usando CXR ainda é uma área em desenvolvimento e é necessário criar e disponibilizar bases de dados de alta qualidade com o mínimo de vieses possível para auxiliar na confiabilidade dos sistemas de IA.

3 MATERIAIS E MÉTODOS

Neste capítulo detalhamos as abordagens experimentais e computacionais utilizadas neste estudo. Isso é fundamental para a compreensão de como os dados foram coletados, processados e analisados. Discutiremos em detalhes os materiais utilizados, incluindo os conjuntos de dados, ferramentas e tecnologias, bem como os métodos de aprendizado profundo e análise dos resultados implementados. Este capítulo fornecerá uma visão abrangente das técnicas e procedimentos empregados, permitindo a reprodutibilidade e validação dos resultados apresentados no estudo. Além disso, quaisquer dificuldades encontradas durante o processo de pesquisa serão destacadas, juntamente com as soluções adotadas para superá-las.

3.1 MODELOS DE VC UTILIZADOS

Neste subcapítulo, descreveremos os modelos de VC utilizados para a classificação de imagens radiográficas e detecção de COVID-19. Os modelos selecionados para este estudo foram a *Resnet50*, o *Vision Transformer* e o *Swin Transformer*. Todos os testes foram realizados sobre implementações na linguagem *Python*, utilizando a biblioteca *PyTorch*. Convém destacar que em todos os experimentos a validação cruzada foi executada em conjuntos de dados distintos dos conjuntos de testes e de treinamentos. As métricas de desempenho e técnicas de interpretação utilizadas para avaliar e comparar os modelos foram: Acurácia, Precisão, Sensibilidade, AUC-ROC e F1-score.

Abaixo, detalharemos cada um desses modelos e a metodologia empregada em sua implementação.

3.1.1 *Convolutional Neural Network* (CNN)

A CNN é um tipo de arquitetura de rede neural profunda especialmente projetada para processar dados de imagem. Ela pode possuir múltiplas camadas de convolução que são capazes de extrair características relevantes de uma imagem através de filtros aplicados localmente. Neste estudo, utilizamos a *Resnet50* para realizar a classificação de imagens radiográficas e a detecção de casos de COVID-19.

No presente experimento para implementar a CNN realizamos pré-processamento nas imagens, aplicando técnicas como redimensionamento, normalização e aumento de dados, para melhorar o desempenho do modelo.

A arquitetura da CNN consiste em várias camadas convolucionais e de *pooling*, seguidas por camadas totalmente conectadas. Utilizamos funções de ativação como *ReLU* e técnicas de regularização, como *dropout*, para evitar overfitting. O modelo foi treinado em um conjunto de treinamento e validação, ajustando os pesos da rede através do algoritmo de *backpropagation* e otimizando a função de perda com o otimizador *Adam*.

A motivação para a escolha do modelo *Resnet50*, dentre tantos outros CNN disponíveis, reside em uma série de fatores. Primeiramente, a *ResNet50* é uma rede neural convolucional profunda com cinquenta camadas, uma evolução significativa em relação aos modelos anteriores com menos camadas. A *ResNet50*, portanto, é capaz de capturar características mais complexas nos dados e, por conseguinte, possui potencial para obter um desempenho superior na classificação de imagens.

Outro aspecto relevante é a característica fundamental da *ResNet50*: a introdução do conceito de conexões residuais, também conhecido como "atalhos" ou "*skip connections*". Esse aspecto permite que o modelo *ResNet50* efetivamente contorne o problema de "desaparecimento do gradiente", um desafio comum encontrado durante o treinamento de redes profundas, facilitando o aprendizado e melhorando a performance do modelo.

Além disso, a arquitetura *ResNet* demonstrou excelentes resultados em diversos problemas de classificação de imagens médicas, o que a torna uma escolha confiável e sólida para tarefas similares. Meedeniya *et al.* (2022) destacaram que essa arquitetura é a que possui a maior tendência de utilização pelos pesquisadores que aplicam modelos CNN ao campo de pesquisa de imagens médicas da atualidade. Sua popularidade também significa que existe uma ampla comunidade de pesquisa em torno dela, com uma abundância de recursos, tutoriais e exemplos de código que podem auxiliar na implementação e na resolução de problemas.

Por fim, muitos *frameworks* modernos de aprendizado de máquina oferecem versões pré-treinadas da *ResNet50*, que podem ser utilizadas como ponto

de partida para transferência de aprendizagem, permitindo economia significativa de tempo e recursos de treinamento.

Baseado em todos estes aspectos, a *ResNet50* se mostrou uma opção altamente atraente e viável para este estudo.

Os hiperparâmetros a serem utilizados para o modelo *ResNet50* estão apresentados na Tabela 5.

Tabela 5 – Parâmetros utilizados no modelo *ResNet50*

Camada	Tipo de Camada	Tamanho do Kernel	Passo (*Stride*)	*Padding*	Saída de Canais
conv1	Conv2d	7x7	2	3	64
bn1	BatchNorm2d	-	-	-	64
relu	ReLU	-	-	-	-
maxpool	MaxPool2d	3x3	2	1	-
layer1_0	BasicBlock/Conv2d	3x3, 3x3	1, 1	1, 1	64, 64, 256
layer1_1	BasicBlock/Conv2d	3x3, 3x3	1, 1	1, 1	64, 64, 256
layer2_0	BasicBlock/Conv2d	3x3, 3x3	2, 1	1, 1	128, 128, 512
layer2_1	BasicBlock/Conv2d	3x3, 3x3	1, 1	1, 1	128, 128, 512
layer3_0	BasicBlock/Conv2d	3x3, 3x3	2, 1	1, 1	256, 256, 1024
layer3_1	BasicBlock/Conv2d	3x3, 3x3	1, 1	1, 1	256, 256, 1024
layer4_0	BasicBlock/Conv2d	3x3, 3x3	2, 1	1, 1	512, 512, 2048
layer4_1	BasicBlock/Conv2d	3x3, 3x3	1, 1	1, 1	512, 512, 2048

Fonte: o autor (2023).

3.1.2 *Vision Transformer* (ViT)

Na implementação do modelo ViT, utilizamos os mesmos conjuntos de dados utilizado para treinar a CNN, com o mesmo pré-processamento aplicado. A arquitetura do modelo consistiu em camadas de *Transformers* empilhadas, com atenção multicabeça e redes de alimentação para processar os *patches* da imagem.

Durante o treinamento do ViT, otimizamos uma função de perda utilizando o otimizador *AdamW*. Realizamos experimentos com diferentes tamanhos de patch, número de camadas de Transformers e hiperparâmetros, a fim de encontrar a configuração ótima para o nosso conjunto de dados. A validação cruzada foi empregada para avaliar o desempenho do modelo em diferentes divisões dos dados.

Quanto ao modelo ViT os parâmetros estão registrados no Quadro 3.

Quadro 3 – Parâmetros utilizados no modelo *Vision Transformer (ViT)*

Camada	Parâmetros
Entrada	Imagem RGB 224 x 224
Incorporação de Patch	Tamanho do patch: 16, Dimensão da incorporação: 768
Codificador Transformer	Número de camadas: 12, Tamanho oculto: 768, Número de cabeças de atenção: 12
Cabeça de Classificação	Tamanho de saída: 2
Pré-treinamento	Sim
Ajuste fino	Apenas última camada

Fonte: o autor (2023).

3.1.3 *Swin Transformer*

Para implementar o *Swin Transformer,* os conjuntos de dados utilizados e o pré-processamento foram os mesmos empregados na CNN e no ViT. A arquitetura do *Swin Transformer* consistiu em camadas de agrupamento de tokens seguidas por camadas de Transformers. Essa abordagem nos permitiu capturar informações contextuais tanto em nível local quanto global.

Durante o treinamento do *Swin Transformer*, aplicamos o otimizador *AdamW* para otimizar a função de perda. Realizamos experimentos para determinar a melhor configuração de hiperparâmetros, como o tamanho dos blocos, o número de camadas de Transformers e as taxas de aprendizado.

Para avaliar o desempenho dos modelos de VC, dividimos o conjunto de dados em conjuntos de treinamento, validação e teste. Utilizamos métricas como acurácia, precisão, sensibilidade e F1-score para avaliar a capacidade de classificação e detecção de COVID-19 dos modelos.

Cada modelo foi treinado e avaliado em um conjunto de dados contendo imagens radiográficas para classificação e detecção da COVID-19.

Por fim os parâmetros do modelo *Swin Transformer* estão demonstrados na Tabela 6.

Tabela 6 – Parâmetros utilizados no modelo *Swin Transformer*

Parâmetros	Valor
Num heads	3, 6, 12, 24
Patch size	4
Depths	2, 2, 6, 6
Embed dim	96
Encoder stride	32

Finetuning task	*Image-classification*
Hidden act	Gelu
Hidden size	768
Layer norm eps	0,00001
Optimize	AdamW
Total optimization steps	8525
Image size	224

Fonte: o autor (2023).

3.2 EQUIPAMENTOS UTILIZADOS

O *hardware* utilizado para executar os experimentos foi um *notebook Dell AlienWare* M15 R7 com CPU Core i7 12700H, 16GB RAM, GPU RTX 3070 8GB VRAM e 1TB SSD e um desktop com CPU AMD A10-7850K, 16GB RAM, GPU RTX 3060 12 GB VRAM e HD de 500 GB.

3.2.1 Datasets utilizados

Em relação ao conjunto de dados foram utilizados nove conjuntos de dados, sendo quatro conjuntos de dados únicos e cinco da aplicação da estratégia CDE e da combinação entre *datasets* contendo Raio X e TC (abordagem híbrida ou de mixagem).

O primeiro foi o COVID-QU-EX *Dataset,* que foi utilizado em Rahman *et al*. (2021), e posteriormente aprimorado por Tahir *et al*. (2022), e está acessível no *Kaggle*[5]. Este conjunto de dados foi criado por uma equipe de pesquisadores da Universidade do Qatar, em Doha, no Catar. Os pesquisadores criaram um banco de dados de imagens de CXR para casos positivos de COVID-19, juntamente com imagens de pneumonia normal e viral. Deste conjunto de dados utilizamos 22.657 imagens, sendo 11.956 de COVID-19 e 10.701 de pulmões saudáveis com resolução de 256x256 *pixels*.

O segundo conjunto de dados foi o HCV-UFPR-COVID-19 (LUZ *et al*., 2021), que consiste em 281 imagens de raio X de pessoas infectadas com COVID-19 e 232 de pessoas que testaram negativo para a doença. Todas as imagens possuem três canais de cor de oito *bits* (RGB), com resolução variando de 2974×2612 a 4248×3480 *pixels*. As imagens são rotuladas em duas classes,

[5] Disponível em: https://www.kaggle.com/datasets/anasmohammedtahir/covidqu

COVID-19 e não COVID, e não há anotações sobre a visualização do ângulo da imagem. O conjunto de dados de raio X HCV-UFPR-COVID-19 é disponibilizado aos pesquisadores mediante solicitação caso a caso. Este conjunto de dados foi criado pelo Hospital da Cruz Vermelha, que recebeu e documentou alguns casos de COVID-19, juntamente com a Universidade Federal do Paraná (UFPR), ambos de Curitiba, Paraná.

O terceiro conjunto de dados foi um conjunto de dados brasileiro, chamado SARS-COV2 CT *dataset* (Mehboob *et al*., 2022; Soares; Angelov, 2020), que está acessível no *Kaggle*[6]. O conjunto de dados SARS CT consiste em 2.481 TC de 120 pacientes, incluindo 1.252 TC de 60 pacientes infectados (*COVID*) e 1.229 TC de 60 pacientes não infectados (*non-COVID*). Os dados foram coletados de pacientes reais de hospitais de São Paulo, Brasil. As imagens de TC variam em tamanho de imagem, pois a menor imagem é 104×153 e a maior imagem é 484×416. O conjunto de dados contém imagens CT heterogêneas com um baixo número de instâncias. Em imagens de CT uma instância refere-se a uma única imagem adquirida durante o exame de CT. Além disso, as imagens de TC têm contraste e resolução diferentes.

O quarto *dataset* utilizado foi o HUST-19, que foi utilizado em Ning *et al*. (2020) e está disponível no *site* da *National Genomics Data Center*[7]. Essa base de dados consiste em 4.001 cortes positivos de TC (pCT) e 9.979 cortes negativos de TC (nCT), selecionados aleatoriamente de 61 pacientes com pneumonia por COVID-19 e 43 pacientes sem pneumonia.

Além dos quatro *datasets* originais, criamos mais dois realizando a mixagem entre os quatro *datasets* (*hybrid*) que foram conforme demonstrados na Tabela 7 abaixo.

[6] Disponível em: https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset.

[7] Disponível em: https://ngdc.cncb.ac.cn/ictcf/HUST-19.php.

Tabela 7 – Estudos e *datasets*

Nº	Fonte	Nome do Dataset	Tipo	Nº total de imagens	Classe COVID-19	Classe Normal
1	Raman *et al*. (2021); Kaggle (2023a)	*COVID-QU-Ex Dataset*	Raio X	22.657	11.956	10.701
2	Luz *et al*. (2021)	HCV-UFPR-COVID-19	Raio X	513	281	232
3	Mehboob *et al*. (2022) e Kaggle (2023b)	*SARS-COV2 CT Dataset*	TC	2.481	1252	1229
4	Ning *et al*. (2020)	HUST-19	TC	13.980	4001	9979
5	O Autor	*DSHybrid1 (COVID-QU-Ex Dataset + HUST-19)*	Raio X + TC	36.637	15.957	20.680
6	O Autor	*DSHybrid2 (HCV-UFPR-COVID-19 + SARS-COV2 CT dataset)*	Raio X + TC	2.994	1.533	1.461
7	O Autor	*CDE Hybrid1&2*	Raio X + TC	39.631	17.490	22.141
8	O Autor	*CDE Raio X (COVID-QU-Ex Dataset& HCV-UFPR-COVID-19)*	Raio X	23.170	12.237	10.933
9	O Autor	CDE TC (HUST-19& SARS-COV2 CT dataset)	TC	16.461	5.253	11.208

Fonte: o autor (2023).

A estratégia adotada teve como objetivo principal avaliar os modelos em um cenário de dados mistos, onde imagens de Raio X e TC foram combinadas. Esta integração levou à utilização da abordagem de CDE para inter-relacionar dois *datasets* criados com essa combinação de dados.

Além disso, empregamos a mesma abordagem do CDE especificamente entre os *datasets* de Raio X e entre aqueles formados apenas por TC. Seguindo

esta metodologia, geramos ainda outros três datasets a partir da interação com os conjuntos híbridos. O propósito dessa abordagem ampla era avaliar a capacidade dos modelos de generalizar as informações presentes tanto em imagens de Raio X quanto em TC, e mais importante, quando essas estavam presentes simultaneamente.

No contexto da classificação de imagens radiográficas, é importante que o *dataset* apresente dados classificados em diferentes categorias. Esse pré-requisito, contudo, representa um desafio significativo. Isso porque, para o treinamento eficaz de um modelo de VC é essencial dispor de milhares de imagens meticulosamente rotuladas.

Para minimizar a carência de grandes datasets será utilizado o recurso de *transfer learning* a partir dos modelos *ResNet50*, ViT e *Swin Transformer* que foram pré-treinados no *dataset ImageNet-1k*.

3.3 *SPLIT* DOS *DATASETS* UTILIZADOS

É importante destacar que todos os nove conjuntos de dados utilizados possuem duas classes (classificação binária) e que os conjuntos de dados utilizados para treinamento, validação e teste dos modelos de VC foram divididos de forma a garantir uma avaliação adequada do desempenho dos modelos. Dessa forma, todos os nove conjuntos de dados foram divididos em treinamento, validação e teste na proporção de 70%, 15% e 15% ou 60%, 20% e 20%, respectivamente. Abaixo, são fornecidos detalhes do split dos datasets.

3.3.1 Conjunto de Dados Original

O conjunto de dados original consiste em imagens radiográficas com anotações de classificação (COVID-19 positivo ou negativo). Essas imagens serão utilizadas como base para a criação dos subconjuntos de treinamento, validação e teste.

3.3.2 Divisão do Conjunto de Dados

Para garantir uma avaliação robusta do desempenho dos modelos, adotaremos a estratégia de divisão tradicionalmente empregada em VC, conhecida como "*train/validation/test split*" (divisão de treinamento, validação e teste).

3.3.2.1 Conjunto de treinamento (*train set*)

O conjunto de treinamento será a parte do conjunto de dados destinada ao treinamento dos modelos. Ele será utilizado para ajustar os pesos das redes neurais e otimizar as funções de perda. Esse conjunto corresponde a uma porcentagem de 70% dos dados disponíveis.

3.3.2.2 Conjunto de validação (*validation* set)

O conjunto de validação será usado para avaliar o desempenho dos modelos durante o treinamento. Ele é utilizado para ajustar hiperparâmetros, como taxas de aprendizado e tamanho do *batch,* e para monitorar o *overfitting*. No trabalho optou-se por adotar a porcentagem de 15% do conjunto de dados original.

3.3.2.3 Conjunto de teste (*test set*)

O conjunto de teste será usado para avaliar o desempenho final dos modelos. Ele é reservado exclusivamente para essa etapa e não é utilizado durante o treinamento ou ajuste de hiperparâmetros. Esse conjunto é fundamental para fornecer uma estimativa imparcial da capacidade de generalização dos modelos. No trabalho optou-se por adotar a porcentagem de 15% ou 20% do conjunto de dados original.

Em resumo, os conjuntos de dados serão divididos em conjuntos de treinamento, validação e teste, respeitando uma proporção adequada. Faremos experimentos com as proporções 70/15/15 e 60/20/20 para treinamento, validação e teste respectivamente.

Após a divisão do conjunto de dados, os modelos serão treinados no conjunto de treinamento, ajustados no conjunto de validação e, finalmente, avaliados no conjunto de teste para estimar sua capacidade de generalização e desempenho em casos não vistos anteriormente.

Essa abordagem de divisão dos *datasets* permite uma avaliação adequada do desempenho dos modelos de VC, garantindo que eles sejam capazes de classificar corretamente as imagens radiográficas e detectar a presença de COVID-19 de forma precisa e confiável. As Tabelas 8 e 9 representam o *split* dos *datasets* para as classes COVID-19 e NORMAL.

Tabela 8 – *Dataset Split* Classe COVID-19

***Dataset Split* Classe COVID-19 (%-70/15/15)**				
Nome do *Dataset*	**COVID-19 (Total)**	**Treinamento**	**Validação**	**Teste**
COVID-QU-Ex	11956	8.369	1793	1794
HCV-UFPR-COVID-19	281	196	43	42
SARS-COV2 CT	1252	876	188	188
HUST-19	4001	2801	600	600
DSHybrid1	15957	11169	2394	2394
DSHybrid2	1533	1073	230	230

Fonte: o autor (2023).

Tabela 9 – *Dataset Split* Classe Normal

***Dataset Split* Classe Normal (%-70/15/15)**				
Nome do *Dataset*	**Normal (Total)**	**Treinamento**	**Validação**	**Teste**
COVID-QU-Ex	10701	7491	1.605	1.605
HCV-UFPR-COVID-19	232	162	35	35
SARS-COV2 CT	1229	860	184	185
HUST-19	9979	6.985	1.497	1497
DSHybrid1	20680	14476	3102	3102
DSHybrid2	1461	1.022	219	220

Fonte: o autor (2023).

3.4 UTILIZAÇÃO DE *CROSS-DATASETS EVALUATION* (CDE)

Diversos estudos têm apontado que os algoritmos de VC não apresentam um desempenho excepcional quando testados em conjuntos de dados nos quais não foram treinados. Para melhorar esse desempenho, pesquisadores têm explorado abordagens de CDE. Zandamela *et al.* (2022) observaram que a utilização de CDE pode auxiliar na melhoria do desempenho de modelos de *DL*. Da mesma forma, Guo *et al.* (2020) destacaram a importância das abordagens de CDE para aprimorar os resultados dos modelos. Além disso, Freire-Obregón *et al.* (2023) defenderam a ideia de que misturar conjuntos de dados durante o treinamento apresenta resultados superiores em comparação com a aplicação

apenas da CDE. Wang *et al*. (2021) defendem que a avaliação entre diferentes conjuntos de dados apoia o seu algoritmo, *OnClass*, como um método robusto para classificação automatizada de tipos de células em conjuntos de dados com grande número de tipos de células não vistas anteriormente.

Todos estes estudos convergem na importância de considerar a variabilidade dos conjuntos de dados, tanto em termos de ambientes de coleta quanto de tipos de dados. Essa consideração é crucial para permitir que o modelo treinado tenha uma melhor capacidade de generalização ao realizar inferências em conjuntos de dados diferentes dos utilizados no treinamento. Portanto, a combinação de técnicas de CDE e a mistura adequada de conjuntos de dados são fundamentais para obter resultados melhores de generalização em diversos contextos e ambientes.

A técnica de CDE desempenha um papel importante na avaliação e validação de modelos de VC. Essa técnica envolve testar e avaliar o desempenho do modelo em conjuntos de dados diferentes dos utilizados durante o treinamento. A técnica proporciona uma estimativa mais abrangente e confiável da capacidade de generalização do modelo, ajudando a verificar se o modelo é capaz de realizar classificações precisas e detecções corretas em dados não vistos anteriormente.

Existem algumas razões pelas quais a CDE é uma prática importante na área de VC:

- generalização do modelo: Ao treinar um modelo em um conjunto de dados específico, ele pode aprender padrões e características únicas desse conjunto em particular. No entanto, é fundamental garantir que o modelo seja capaz de generalizar e realizar classificações precisas em novos conjuntos de dados. A avaliação cruzada permite testar o desempenho do modelo em diferentes conjuntos de dados e verificar sua capacidade de adaptação a diferentes contextos.
- variedade de dados: conjuntos de dados diferentes podem conter variações nas condições de captura das imagens, como iluminação, ângulo de visão, qualidade da imagem e outros fatores. A avaliação em diferentes conjuntos de dados permite que o modelo seja testado em uma variedade de cenários, fornecendo uma avaliação mais completa de sua robustez e confiabilidade.

- detecção de viés e sobreajuste (*overfitting):* a avaliação cruzada também ajuda a identificar possíveis problemas de viés ou sobreajuste do modelo. Ao testar o modelo em conjuntos de dados independentes, é possível verificar se o modelo está aprendendo características específicas do conjunto de treinamento que não se aplicam a outras imagens ou domínios. Isso ajuda a evitar o sobreajuste do modelo em relação a determinados padrões e a garantir uma generalização mais equilibrada.
- comparação de desempenho: através da avaliação cruzada, é possível comparar o desempenho de diferentes modelos em vários conjuntos de dados. Isso permite avaliar quais modelos são mais robustos e consistentes em diferentes cenários, auxiliando na seleção do melhor modelo para uma determinada aplicação.

Em suma, a CDE é uma técnica importante para garantir a qualidade e a confiabilidade dos modelos de VC. Ela permite uma avaliação mais abrangente, ajudando a identificar possíveis problemas de generalização, viés e sobreajuste além de fornecer uma base sólida para comparação de desempenho entre diferentes modelos.

Cabe destacar que, para os *datasets* utilizados na aplicação da técnica de CDE, foram adotadas proporções ligeiramente diferentes para *split dataset* para satisfazer a restrição de que os dados utilizados no conjunto de teste fossem oriundos de *datasets* distintos do que foi utilizado no treinamento e na validação. Dessa forma, não faz muito sentido considerar os dados de teste como um percentual do outro conjunto de dados, uma vez que se trata de outro conjunto de dados que não é parte do conjunto utilizado no treinamento e validação.

Para a estratégia de CDE o *data split* da abordagem nas imagens de Raio X foi realizado como demonstrado nas Tabelas 10 e 11 abaixo:

Tabela 10 – CDE Raio X *Dataset Split* Classe COVID-19

CDE Raio X *Dataset Split* Classe COVID-19				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
COVID-QU-Ex	11956	8.369	3.587	0
HCV-UFPR-COVID-19	281	0	0	281

Fonte: o autor (2023).

Tabela 11 – CDE Raio X *Dataset Split* Classe Normal

CDE Raio X *Dataset Split* Classe Normal				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
COVID-QU-Ex	10701	7491	3.210	0
HCV-UFPR-COVID-19	232	0	0	232

Fonte: o autor (2023).

Para a estratégia de CDE o *data split* da abordagem nas imagens de TC foi realizado como demonstrado nas Tabelas 12 e 13 abaixo:

Tabela 12 – CDE TC *Dataset Split* Classe COVID-19

CDE TC *Dataset Split* Classe COVID-19				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
HUST-19	4001	2801	1200	0
SARS-COV2 CT	1252	0	0	1252

Fonte: o autor (2023).

Tabela 13 – CDE TC *Dataset Split* Classe Normal

CDE TC *Dataset Split* Classe Normal				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
HUST-19	9979	6.985	2.994	0
SARS-COV2 CT	1229	0	0	1229

Fonte: o autor (2023).

Para a estratégia de CDE, o *data split* da abordagem híbrida foi realizado como demonstrado nas Tabelas 14 e 15 abaixo:

Tabela 14 – CDE *Hybrid1&2 Dataset Split* Classe COVID-19

CDE Hybrid1&2 *Dataset Split* Classe COVID-19				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
DSHybrid1	15957	11169	4.788	0
DSHybrid2	1533	0	0	1533

Fonte: o autor (2023).

Tabela 15 – CDE *Hybrid1&2 Dataset Split* Classe Normal

CDE Hybrid1&2 *Dataset Split* Classe Normal				
Nome do *Dataset*	**Normal**	**Treinamento (70%)**	**Validação (30%)**	**Teste (100%)**
DSHybrid1	20680	14476	6204	0
DSHybrid2	1461	0	0	1461

Fonte: o autor (2023).

3.5 AJUSTE DE HIPERPARÂMETROS COM A BIBLIOTECA *OPTUNA*

Com o objetivo de otimizar as métricas que são utilizadas em visão computacional uma técnica essencial é o uso da otimização de hiperparâmetros. Como destacado por alguns autores, quando otimizados e aplicados aos modelos, os hiperparâmetros podem melhorar significativamente as métricas. Akiba *et al*. (2019) apresentaram o *Optuna*, um *framework* de otimização de hiperparâmetros fundamentado na inovadora abordagem "*define-by-run*". Esta pesquisa delineou princípios fundamentais para futuros *softwares* de otimização, destacando a importância da adaptabilidade, eficiência e *design* multifacetado. O *Optuna* se destaca pela configuração dinâmica do espaço de busca de parâmetros, mecanismos refinados de busca e poda, e uma arquitetura flexível que se integra perfeitamente em um espectro de ambientes, desde sistemas distribuídos expansivos até configurações experimentais mais focadas. Superando as limitações tradicionais das metodologias de ajuste de hiperparâmetros, o princípio *define-by-run* do *Optuna* representa uma evolução paradigmática no campo. A robustez do sistema é enfatizada através de sua fusão de técnicas de busca e poda e sua adaptabilidade, atendendo às necessidades de uma base de usuários diversificada e propósitos. Embora o artigo culmine com uma perspectiva otimista sobre a trajetória futura dos *frameworks* de otimização, fundamentando sua confiança no *Optuna*, ele também faz um chamado para uma análise comparativa aprofundada contra ferramentas existentes para reforçar suas afirmações empíricas. Fundamentalmente, o trabalho de Akiba *et al*. (2019) atua como um elemento central na otimização de hiperparâmetros, abrindo novos caminhos e estimulando avanços futuros na disciplina.

Shekhar *et al*. (2021) se aprofundam no domínio da otimização de hiperparâmetros em *machine learning*, destacando as limitações dos métodos tradicionais e voltando sua atenção para alternativas avançadas, focando especialmente em bibliotecas *Python* como *Optuna*, *HyperOpt*, *Optunity* e SMAC. Através de avaliações em seis conjuntos de dados do mundo real, eles destacaram o desempenho superior do Optuna para seleção de algoritmos e tarefas de hiperparâmetros e a eficácia do *HyperOpt* para o desafio do *perceptron* multicamada. Apesar dos substanciais contribuições do estudo, ele poderia se beneficiar de uma justificativa mais explícita por trás da seleção de ferramentas, *in-*

sights metodológicos expandidos e discussões mais aprofundadas sobre compromissos de desempenho e limitações potenciais. Shekhar *et al*. (2021) oferecem uma análise comparativa fundamental em ferramentas otimização de hiperparâmetros, mas indica áreas propícias para uma exploração mais profunda e clareza.

4 EXPERIMENTOS, RESULTADOS E DISCUSSÃO

Todos os modelos executaram classificação binária sobre as classes COVID-19 e NORMAL (Classificação Binária). O treinamento foi executado por 25 épocas em cada um dos três modelos (*ResNet50*, ViT e *SwinT*) nos nove conjuntos de dados sem utilização de otimização. Mais tarde, foi realizada otimização bayesiana com a biblioteca *Optuna* onde foram encontrados hiperparâmetros otimizados para CNN, ViT e *Swin Transformer*. Adotamos, ainda, diferentes *split datasets* (70/15/15 e 60/20/20) para as execuções com os hiperparâmetros otimizados para os otimizadores *AdamW* e *Stochastic Gradient Descent* (SGD). Os hiperparâmetros de cada otimizador e modelo será descrito em cada subcapítulo subsequente.

Dessa forma, foram obtidos 675 resultados compostos pelas métricas AUC, acurácia, precisão, sensibilidade e F1-Score. Isso foi possível porque foram nove *datasets* multiplicados por cinco métricas multiplicadas por cinco estratégias distintas de *split datasets* e otimizadores multiplicados pelos três modelos (CNN, ViT e SwinT) $(9\text{x}5\text{x}5\text{x}3 = 675)$.

Adicionalmente, foi plotada a matriz de confusão, a curva ROC, o relatório de classificação e um *grid* contendo 16 imagens com os respectivos rótulos verdadeiros e preditos em cada modelo e *dataset.*

4.1 EXPERIMENTO 1: APLICAÇÃO DO MODELO DE REDES NEURAIS CONVOLUCIONAIS *RESNET50*

4.1.1 Resultados do Experimento 1

Para facilitar a análise dos resultados obtidos nos nove *datasets* utilizados, os dados foram compilados na Tabela 16.

Tabela 16 – Resultados obtidos com CNN nos nove *datasets* sem otimização e com *split dataset* 70/15/15

Modelo	Dataset	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0,99	0,94	0,94	0,94	0,94
	HCV-UFPR-COVID-19	Raio X	0,75	0,7	0,7	0,7	0,7

	HUST-19	TC	0,99	0,96	0,96	0,96	0,96
	SARS-COV-2 Ct	TC	0,98	0,89	0,9	0,89	0,89
	DSHybrid1	Raio X + TC	0,99	0,92	0,92	0,92	0,92
	DSHybrid2	Raio X + TC	0,94	0,87	0,87	0,87	0,87
	CDE Hybrid1&2	Raio X + TC	0,72	0,65	0,67	0,65	0,65
	CDE Raio X	Raio X	0,53	0,55	0,54	0,55	0,41
	CDE TC	TC	0,79	0,6	0,74	0,6	0,54

Fonte: o autor (2023).

Após a execução do CNN sem otimização de hiperparâmetros, foi realizada uma otimização bayesiana que resultou em hiperparâmetros otimizados para o modelo, cujos valores foram compilados e apresentados na Tabela 17:

Tabela 17 – Hiperparâmetros otimizados para o modelo CNN

Optimizer	***Learning Rate***	***Weight Decay***	***Batch Size***	**Momentum**
Adam	0.002931	-	-	-
SGD	0.000767	0.000004	32	0.173

Fonte: o autor (2023).

A Tabelo 18 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador Adam:

Tabela 18 – Resultados obtidos com CNN nos nove *datasets* com otimizador Adam e *split dataset* 70/15/15

Modelo	**Dataset**	**Tipo de Imagem**	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**
CNN	COVID-QU-Ext	Raio X	0.98	0.93	0.93	0.93	0.93
	HCV-UFPR-COVID-19	Raio X	0.89	0.78	0.78	0.73	0.74
	HUST-19	TC	1.0	0.99	0.99	0.99	0.99
	SARS-COV-2 Ct	TC	0.97	0.9	0.9	0.9	0.9
	DSHybrid1	Raio X + TC	0.99	0.94	0.95	0.93	0.94
	DSHybrid2	Raio X + TC	0.92	0.85	0.85	0.85	0.85
	CDE Hybrid1&2	Raio X + TC	0.7	0.65	0.67	0.65	0.64
	CDE Raio X	Raio X	0.55	0.55	0.52	0.5	0.39
	CDE TC	TC	0.78	0.64	0.75	0.65	0.61

Fonte: o autor (2023).

A Tabela 19 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 19 – Resultados obtidos com CNN nos nove *datasets* com otimizador SGD e *split dataset* 70/15/15

Modelo	Dataset	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.96	0.89	0.89	0.89	0.89
	HCV-UFPR-COVID-19	Raio X	0.91	0.77	0.81	0.77	0.76
	HUST-19	TC	0.99	0.98	0.97	0.97	0.97
	SARS-COV-2 Ct	TC	0.95	0.87	0.88	0.88	0.87
	DSHybrid1	Raio X + TC	0.99	0.94	0.94	0.94	0.94
	DSHybrid2	Raio X + TC	0.91	0.84	0.84	0.84	0.84
	CDE Hybrid1&2	Raio X + TC	0.67	0.61	0.63	0.61	0.59
	CDE Raio X	Raio X	0.5	0.51	0.41	0.47	0.38
	CDE TC	TC	0.78	0.62	0.74	0.62	0.57

Fonte: o autor (2023).

A Tabela 20 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador Adam:

Tabela 20 – Resultados obtidos com CNN nos nove *datasets* com otimizador Adam e *split dataset* 60/20/20

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.98	0.92	0.92	0.91	0.91
	HCV-UFPR-COVID-19	Raio X	0.9	0.81	0.81	0.81	0.81
	HUST-19	TC	1.0	0.99	0.99	0.99	0.99
	SARS-COV-2 Ct	TC	0.96	0.89	0.9	0.89	0.89
	DSHybrid1	Raio X + TC	0.99	0.92	0.92	0.93	0.92
	DSHybrid2	Raio X + TC	0.95	0.86	0.86	0.86	0.86
	CDE Hybrid1&2	Raio X + TC	0.69	0.63	0.65	0.64	0.63
	CDE Raio X	Raio X	0.48	0.45	0.37	0.42	0.38
	CDE TC	TC	0.78	0.62	0.74	0.62	0.69

Fonte: o autor (2023).

A Tabela 21 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 21 – Resultados obtidos com CNN nos nove *datasets* com otimizador SGD e split *dataset* 60/20/20

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.96	0.9	0.9	0.9	0.9
	HCV-UFPR-COVID-19	Raio X	0.84	0.77	0.76	0.75	0.76

	HUST-19	TC	0.99	0.97	0.97	0.96	0.97
	SARS-COV-2 Ct	TC	0.93	0.86	0.87	0.87	0.86
	DSHybrid1	Raio X + TC	0.97	0.91	0.91	0.91	0.91
	DSHybrid2	Raio X + TC	0.91	0.82	0.83	0.82	0.82
	CDE Hybrid1&2	Raio X + TC	0.67	0.61	0.63	0.61	0.62
	CDE Raio X	Raio X	0.46	0.54	0.47	0.49	0.48
	CDE TC	TC	0.77	0.57	0.72	0.58	0.64

Fonte: o autor (2023).

4.2 EXPERIMENTO 2: APLICAÇÃO DO MODELO *VISION TRANSFORMER* (VIT)

4.2.1 Resultados do Experimento 2

Para facilitar a análise dos resultados obtidos em nove *datasets* utilizados, os dados foram compilados na Tabela 22.

Tabela 22 – Resultados obtidos com ViT nos nove *datasets* sem otimização e com *split dataset* 70/15/15

Modelo	***Dataset***	**Tipo de imagem**	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**
ViT	COVID-QU-Ext	Raio X	0,99	0,95	0,95	0,95	0,95
	HCV-UFPR-COVID-19	Raio X	0,76	0,7	0,7	0,7	0,7
	HUST-19	TC	0,99	0,94	0,94	0,94	0,94
	SARS-COV-2 Ct	TC	0,99	0,95	0,95	0,95	0,95
	DSHybrid1	Raio X + TC	0,98	0,93	0,93	0,93	0,93
	DSHybrid2	Raio X + TC	0,96	0,89	0,9	0,89	0,89
	CDE Hybrid1&2	Raio X + TC	0,67	0,61	0,62	0,61	0,61
	CDE Raio X	Raio X	0,52	0,52	0,46	0,52	0,43
	CDE TC	TC	0,28	0,61	0,73	0,61	0,56

Fonte: o autor (2023).

Após a execução do ViT sem otimização de hiperparâmetros, foi realizada uma otimização bayesiana que resultou em hiperparâmetros otimizados para o modelo, cujos valores foram os seguintes:

Tabela 23 – Hiperparâmetros otimizados para o modelo ViT

Optimizer	***Learning Rate***	***Weight Decay***	***Batch Size***	***Momentum***	***Beta 1***	***Beta2***
AdamW	0.000015	0.000029	16	-	0.851	0.995
SGD	0.000847	0.000248	8	0.866	-	-

Fonte: o autor (2023).

A Tabela 24 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador AdamW:

Tabela 24 – Resultados obtidos com ViT nos nove *datasets* com otimizador AdamW e *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
ViT	COVID-QU-Ext	Raio X	0.93	0.93	0.93	0.93	0.93
	HCV-UFPR-COVID-19	Raio X	0.93	0.62	0.63	0.63	0.62
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.76	0.74	0.66	0.62	0.95
	DSHybrid1	Raio X + TC	0.93	0.93	0.94	0.93	0.93
	DSHybrid2	Raio X + TC	0.84	0.84	0.84	0.84	0.84
	CDE Hybrid1&2	Raio X + TC	0.63	0.63	0.64	0.63	0.63
	CDE Raio X	Raio X	0.52	0.56	0.59	0.52	0.42
	CDE TC	TC	0.7	0.7	0.71	0.7	0.69

Fonte: o autor (2023).

A Tabela 25 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 25 – Resultados obtidos com ViT nos nove *datasets* com otimizador SDG e *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
ViT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.98	0.79	0.81	0.78	0.79
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.74	0.79	0.77	0.62	0.67
	DSHybrid1	Raio X + TC	0.98	0.99	0.98	0.98	0.98
	DSHybrid2	Raio X + TC	0.95	0.95	0.95	0.95	0.95
	CDE Hybrid1&2	Raio X + TC	0.72	0.72	0.72	0.72	0.72
	CDE Raio X	Raio X	0.5	0.55	0.27	0.5	0.35
	CDE TC	TC	0.76	0.77	0.76	0.76	0.76

Fonte: o autor (2023).

A Tabela 26 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador AdamW:

Tabela 26 – Resultados obtidos com ViT nos nove *datasets* com otimizador AdamW e *split dataset* 60/20/20

Modelo	*Dataset*	Tipo de imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
ViT	COVID-QU-Ext	Raio X	0.91	0.91	0.91	0.91	0.91
	HCV-UFPR-COVID-19	Raio X	0.66	0.64	0.67	0.64	0.66
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.73	0.8	0.75	0.59	0.68
	DSHybrid1	Raio X + TC	0.94	0.95	0.95	0.94	0.94
	DSHybrid2	Raio X + TC	0.88	0.88	0.88	0.88	0.88
	CDE Hybrid1&2	Raio X + TC	0.63	0.63	0.63	0.63	0.63
	CDE Raio X	Raio X	0.5	0.54	0.5	0.5	0.51
	CDE TC	TC	0.69	0.69	0.73	0.69	0.72

Fonte: o autor (2023).

A Tabela 27 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 27 – Resultados obtidos com ViT nos nove *datasets* com otimizador SGD e *split dataset* 60/20/20

Modelo	*Dataset*	Tipo de imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
ViT	COVID-QU-Ext	Raio X	0.97	0.97	0.97	0.97	0.97
	HCV-UFPR-COVID-19	Raio X	0.85	0.85	0.86	0.85	0.85
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.82	0.85	0.84	0.7	0.76
	DSHybrid1	Raio X + TC	0.98	0.98	0.98	0.98	0.98
	DSHybrid2	Raio X + TC	0.93	0.93	0.93	0.93	0.93
	CDE Hybrid1&2	Raio X + TC	0.71	0.71	0.71	0.71	0.71
	CDE Raio X	Raio X	0.5	0.55	0.44	0.5	0.5
	CDE TC	TC	0.72	0.73	0.73	0.72	0.73

Fonte: o autor (2023).

4.3 EXPERIMENTO 3: APLICAÇÃO DO MODELO *SWIN TRANSFORMER*

4.3.1 Resultados do Experimento 3

Para facilitar a análise dos resultados obtidos em nove *datasets* utilizados, os dados foram compilados na Tabela 28.

Tabela 28 – Resultados obtidos com SwinT nos nove *datasets* sem otimização e com *split dataset 70*/15/15

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
SwinT	COVID-QU-Ext	Raio X	0,98	0,98	0,98	0,98	0,98
	HCV-UFPR-COVID-19	Raio X	0,72	0,71	0,73	0,71	0,71
	HUST-19	TC	0,99	0,99	0,99	0,99	0,99
	SARS-COV-2 Ct	TC	0,98	0,98	0,98	0,98	0,98
	DSHybrid1	Raio X + TC	0,99	0,99	0,99	0,99	0,99
	DSHybrid2	Raio X + TC	0,96	0,96	0,96	0,96	0,96
	CDE Hybrid1&2	Raio X + TC	0,74	0,74	0,74	0,74	0,74
	CDE Raio X	Raio X	0,5	0,54	0,45	0,54	0,39
	CDE TC	TC	0,73	0,73	0,73	0,73	0,73

Fonte: o autor (2023).

Após a execução do *Swin Transformer* sem otimização de hiperparâmetros, foi realizada uma otimização bayesiana que resultou em hiperparâmetros otimizados para o modelo, cujos valores foram os seguintes:

Tabela 29 – Hiperparâmetros otimizados para o modelo SwinT

Optimizer	*Learning Rate*	*Weight Decay*	*Batch Size*	*Momentum*	Beta 1	Beta2
AdamW	0.000015	0.000029	16	-	0.851	0.995
SGD	0.000847	0.000248	8	0.866	-	-

Fonte: o autor (2023).

A Tabela 30 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador AdamW:

Tabela 30 – Resultados obtidos com SwinT nos nove *datasets* sem otimizador AdamW e *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
SwinT	COVID-QU-Ext	Raio X	0.99	0.99	0.99	0.99	0.99
	HCV-UFPR-COVID-19	Raio X	0.78	0.79	0.83	0.78	0.78
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97

	DSHybrid1	Raio X + TC	0.98	0.98	0.98	0.98	0.98
	DSHybrid2	Raio X + TC	0.97	0.97	0.97	0.97	0.97
	CDE Hybrid1&2	Raio X + TC	0.73	0.73	0.73	0.73	0.73
	CDE Raio X	Raio X	0.51	0.56	0.71	0.51	0.38
	CDE TC	TC	0.69	0.69	0.7	0.69	0.69

Fonte: o autor (2023).

A Tabela 31 apresenta os resultados obtidos com o *split dataset* 70/15/15 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 31 – Resultados obtidos com SwinT nos nove *datasets* sem otimizador SGD e *split dataset 70*/15/15

Modelo	***Dataset***	**Tipo de Imagem**	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**
SwinT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.87	0.86	0.87	0.87	0.86
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97
	DSHybrid1	Raio X + TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X + TC	0.91	0.91	0.92	0.91	0.91
	CDE Hybrid1&2	Raio X + TC	0.75	0.75	0.75	0.75	0.75
	CDE Raio X	Raio X	0.5	0.55	0.27	0.5	0.35
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

A Tabela 32 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador AdamW:

Tabela 32 – Resultados obtidos com SwinT nos nove *datasets* sem otimizador AdamW e *split dataset 60/20/20*

Modelo	***Dataset***	**Tipo de Imagem**	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**
SwinT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.94	0.94	0.94	0.94	0.94
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0

	SARS-COV-2 Ct	TC	0.94	0.94	0.94	0.94	0.94
	DSHybrid1	Raio X + TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X + TC	0.94	0.94	0.94	0.94	0.94
	CDE Hybrid1&2	Raio X + TC	0.74	0.74	0.75	0.74	0.74
	CDE Raio X	Raio X	0.5	0.55	0.61	0.5	0.48
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

A Tabela 33 apresenta os resultados obtidos com o *split dataset* 60/20/20 com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 33 – Resultados obtidos com SwinT nos nove *datasets* sem otimizador SGD e *split dataset* 60/20/20

Modelo	***Dataset***	**Tipo de Imagem**	**AUC**	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**
SwinT	COVID-QU-Ext	Raio X	0.97	0.97	0.97	0.97	0.97
	HCV-UFPR-COVID-19	Raio X	0.77	0.73	0.77	0.77	0.77
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97
	DSHybrid1	Raio X + TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X + TC	0.95	0.95	0.95	0.95	0.95
	CDE Hybrid1&2	Raio X + TC	0.71	0.71	0.71	0.71	0.71
	CDE Raio X	Raio X	0.5	0.55	0.77	0.5	0.5
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

4.4 ANÁLISE ESTATÍSTICA E COMPARAÇÕES ENTRE OS RESULTADOS OBTIDOS NOS TRÊS EXPERIMENTOS

Para facilitar a análise estatística e a comparação dos resultados entre as diferentes estratégias de divisão de conjunto de dados e otimizadores, foram criadas as cinco Tabelas (34, 35, 36, 37 e 38), que compilam um total de 675 valores de métricas agregadas para os três modelos executados nos nove con-

juntos de dados diferentes, de acordo com as estratégias de divisão de conjunto de dados e o uso de otimizadores correspondentes.

4.4.1 Agregação dos Resultados dos Experimento Segundo a Estratégia de *Split Dataset* e Emprego de Otimizadores

A Tabela 34 apresenta os resultados obtidos com o *split dataset* 70/15/15 sem otimização:

Tabela 34 – Grupo de métricas agregadas para os três modelos executados sem otimização e com *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0,99	0,94	0,94	0,94	0,94
	HCV-UFPR-COVID-19	Raio X	0,75	0,7	0,7	0,7	0,7
	HUST-19	TC	0,99	0,96	0,96	0,96	0,96
	SARS-COV-2 Ct	TC	0,98	0,89	0,9	0,89	0,89
	DSHybrid1	Raio X+TC	0,99	0,92	0,92	0,92	0,92
	DSHybrid2	Raio X+TC	0,94	0,87	0,87	0,87	0,87
	CDE Hybrid1&2	Raio X+TC	0,72	0,65	0,67	0,65	0,65
	CDE Raio X	Raio X	0,53	0,55	0,54	0,55	0,41
	CDE TC	TC	0,79	0,6	0,74	0,6	0,54
ViT	COVID-QU-Ext	Raio X	0,99	0,95	0,95	0,95	0,95
	HCV-UFPR-COVID-19	Raio X	0,76	0,7	0,7	0,7	0,7
	HUST-19	TC	0,99	0,94	0,94	0,94	0,94
	SARS-COV-2 Ct	TC	0,99	0,95	0,95	0,95	0,95
	DSHybrid1	Raio X+TC	0,98	0,93	0,93	0,93	0,93
	DSHybrid2	Raio X+TC	0,96	0,89	0,9	0,89	0,89
	CDE Hybrid1&2	Raio X+TC	0,67	0,61	0,62	0,61	0,61
	CDE Raio X	Raio X	0,52	0,52	0,46	0,52	0,43
	CDE TC	TC	0,28	0,61	0,73	0,61	0,56
SwinT	COVID-QU-Ext	Raio X	0,98	0,98	0,98	0,98	0,98
	HCV-UFPR-COVID-19	Raio X	0,72	0,71	0,73	0,71	0,71
	HUST-19	TC	0,99	0,99	0,99	0,99	0,99
	SARS-COV-2 Ct	TC	0,98	0,98	0,98	0,98	0,98
	DSHybrid1	Raio X+TC	0,99	0,99	0,99	0,99	0,99
	DSHybrid2	Raio X+TC	0,96	0,96	0,96	0,96	0,96
	CDE Hybrid1&2	Raio X+TC	0,74	0,74	0,74	0,74	0,74
	CDE Raio X	Raio X	0,5	0,54	0,45	0,54	0,39
	CDE TC	TC	0,73	0,73	0,73	0,73	0,73

Fonte: o autor (2023).

A Tabela 35 apresenta os resultados obtidos com *split dataset* 70/15/15 e com os hiperparâmetros otimizados para o otimizador Adam/AdamW:

Tabela 35 – Grupo de métricas agregadas para os três modelos executados com otimizador Adam/AdamW e *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibi-lidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.98	0.93	0.93	0.93	0.93
	HCV-UFPR-COVID-19	Raio X	0.89	0.78	0.78	0.73	0.74
	HUST-19	TC	1.0	0.99	0.99	0.99	0.99
	SARS-COV-2 Ct	TC	0.97	0.9	0.9	0.9	0.9
	DSHybrid1	Raio X+TC	0.99	0.94	0.95	0.93	0.94
	DSHybrid2	Raio X+TC	0.92	0.85	0.85	0.85	0.85
	CDE Hybrid1&2	Raio X+TC	0.7	0.65	0.67	0.65	0.64
	CDE Raio X	Raio X	0.55	0.55	0.52	0.5	0.39
	CDE TC	TC	0.78	0.64	0.75	0.65	0.61
ViT	COVID-QU-Ext	Raio X	0.93	0.93	0.93	0.93	0.93
	HCV-UFPR-COVID-19	Raio X	0.93	0.62	0.63	0.63	0.62
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.76	0.74	0.66	0.62	0.95
	DSHybrid1	Raio X+TC	0.93	0.93	0.94	0.93	0.93
	DSHybrid2	Raio X+TC	0.84	0.84	0.84	0.84	0.84
	CDE Hybrid1&2	Raio X+TC	0.63	0.63	0.64	0.63	0.63
	CDE Raio X	Raio X	0.52	0.56	0.59	0.52	0.42
	CDE TC	TC	0.7	0.7	0.71	0.7	0.69
SwinT	COVID-QU-Ext	Raio X	0.99	0.99	0.99	0.99	0.99
	HCV-UFPR-COVID-19	Raio X	0.78	0.79	0.83	0.78	0.78
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97
	DSHybrid1	Raio X+TC	0.98	0.98	0.98	0.98	0.98
	DSHybrid2	Raio X+TC	0.97	0.97	0.97	0.97	0.97
	CDE Hybrid1&2	Raio X+TC	0.73	0.73	0.73	0.73	0.73
	CDE Raio X	Raio X	0.51	0.56	0.71	0.51	0.38
	CDE TC	TC	0.69	0.69	0.7	0.69	0.69

Fonte: o autor (2023).

A Tabela 36 apresenta os resultados obtidos com *split dataset* 70/15/15 e com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 36 – Grupo de métricas agregadas para os três modelos executados com otimizador SGD e *split dataset* 70/15/15

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.96	0.89	0.89	0.89	0.89
	HCV-UFPR-COVID-19	Raio X	0.91	0.77	0.81	0.77	0.76
	HUST-19	TC	0.99	0.98	0.97	0.97	0.97
	SARS-COV-2 Ct	TC	0.95	0.87	0.88	0.88	0.87
	DSHybrid1	Raio X+TC	0.99	0.94	0.94	0.94	0.94
	DSHybrid2	Raio X+TC	0.91	0.84	0.84	0.84	0.84
	CDE Hybrid1&2	Raio X+TC	0.67	0.61	0.63	0.61	0.59
	CDE Raio X	Raio X	0.5	0.51	0.41	0.47	0.38
	CDE TC	TC	0.78	0.62	0.74	0.62	0.57
ViT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.98	0.79	0.81	0.78	0.79
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.74	0.79	0.77	0.62	0.67
	DSHybrid1	Raio X+TC	0.98	0.99	0.98	0.98	0.98
	DSHybrid2	Raio X+TC	0.95	0.95	0.95	0.95	0.95
	CDE Hybrid1&2	Raio X+TC	0.72	0.72	0.72	0.72	0.72
	CDE Raio X	Raio X	0.5	0.55	0.27	0.5	0.35
	CDE TC	TC	0.76	0.77	0.76	0.76	0.76
SwinT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.87	0.86	0.87	0.87	0.86
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97
	DSHybrid1	Raio X+TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X+TC	0.91	0.91	0.92	0.91	0.91
	CDE Hybrid1&2	Raio X+TC	0.75	0.75	0.75	0.75	0.75
	CDE Raio X	Raio X	0.5	0.55	0.27	0.5	0.35
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

A Tabela 37 apresenta os resultados obtidos com *split dataset* 60/20/20 e com os hiperparâmetros otimizados para o otimizador Adam/AdamW:

Tabela 37 – Grupo de métricas agregadas para os três modelos executados com otimizador Adam/AdamW e com *split dataset* 60/20/20

Modelo	*Dataset*	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibilidade	F1-Score

CNN	COVID-QU-Ext	Raio X	0.98	0.92	0.92	0.91	0.91
	HCV-UFPR-COVID-19	Raio X	0.9	0.81	0.81	0.81	0.81
	HUST-19	TC	1.0	0.99	0.99	0.99	0.99
	SARS-COV-2 Ct	TC	0.96	0.89	0.9	0.89	0.89
	DSHybrid1	Raio X+TC	0.99	0.92	0.92	0.93	0.92
	DSHybrid2	Raio X+TC	0.95	0.86	0.86	0.86	0.86
	CDE Hybrid1&2	Raio X+TC	0.69	0.63	0.65	0.64	0.63
	CDE Raio X	Raio X	0.48	0.45	0.37	0.42	0.38
	CDE TC	TC	0.78	0.62	0.74	0.62	0.69
ViT	COVID-QU-Ext	Raio X	0.91	0.91	0.91	0.91	0.91
	HCV-UFPR-COVID-19	Raio X	0.66	0.64	0.67	0.64	0.66
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.73	0.8	0.75	0.59	0.68
	DSHybrid1	Raio X+TC	0.94	0.95	0.95	0.94	0.94
	DSHybrid2	Raio X+TC	0.88	0.88	0.88	0.88	0.88
	CDE Hybrid1&2	Raio X+TC	0.63	0.63	0.63	0.63	0.63
	CDE Raio X	Raio X	0.5	0.54	0.5	0.5	0.51
	CDE TC	TC	0.69	0.69	0.73	0.69	0.72
SwinT	COVID-QU-Ext	Raio X	0.98	0.98	0.98	0.98	0.98
	HCV-UFPR-COVID-19	Raio X	0.94	0.94	0.94	0.94	0.94
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.94	0.94	0.94	0.94	0.94
	DSHybrid1	Raio X+TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X+TC	0.94	0.94	0.94	0.94	0.94
	CDE Hybrid1&2	Raio X+TC	0.74	0.74	0.75	0.74	0.74
	CDE Raio X	Raio X	0.5	0.55	0.61	0.5	0.48
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

A Tabela 38 apresenta os resultados obtidos com *split dataset* 60/20/20 e com os hiperparâmetros otimizados para o otimizador SGD:

Tabela 38 – Grupo de métricas agregadas para os três modelos executados com otimizador SGD e *split dataset* 60/20/20

Modelo	Dataset	Tipo de Imagem	AUC	Acurácia	Precisão	Sensibili-dade	F1-Score
CNN	COVID-QU-Ext	Raio X	0.96	0.9	0.9	0.9	0.9
	HCV-UFPR-COVID-19	Raio X	0.84	0.77	0.76	0.75	0.76
	HUST-19	TC	0.99	0.97	0.97	0.96	0.97
	SARS-COV-2 Ct	TC	0.93	0.86	0.87	0.87	0.86

	DSHybrid1	Raio X+TC	0.97	0.91	0.91	0.91	0.91
	DSHybrid2	Raio X+TC	0.91	0.82	0.83	0.82	0.82
	CDE Hybrid1&2	Raio X+TC	0.67	0.61	0.63	0.61	0.62
	CDE Raio X	Raio X	0.46	0.54	0.47	0.49	0.48
	CDE TC	TC	0.77	0.57	0.72	0.58	0.64
ViT	COVID-QU-Ext	Raio X	0.97	0.97	0.97	0.97	0.97
	HCV-UFPR-COVID-19	Raio X	0.85	0.85	0.86	0.85	0.85
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.82	0.85	0.84	0.7	0.76
	DSHybrid1	Raio X+TC	0.98	0.98	0.98	0.98	0.98
	DSHybrid2	Raio X+TC	0.93	0.93	0.93	0.93	0.93
	CDE Hybrid1&2	Raio X+TC	0.71	0.71	0.71	0.71	0.71
	CDE Raio X	Raio X	0.5	0.55	0.44	0.5	0.5
	CDE TC	TC	0.72	0.73	0.73	0.72	0.73
SwinT	COVID-QU-Ext	Raio X	0.97	0.97	0.97	0.97	0.97
	HCV-UFPR-COVID-19	Raio X	0.77	0.73	0.77	0.77	0.77
	HUST-19	TC	1.0	1.0	1.0	1.0	1.0
	SARS-COV-2 Ct	TC	0.97	0.97	0.97	0.97	0.97
	DSHybrid1	Raio X+TC	0.99	0.99	0.99	0.99	0.99
	DSHybrid2	Raio X+TC	0.95	0.95	0.95	0.95	0.95
	CDE Hybrid1&2	Raio X+TC	0.71	0.71	0.71	0.71	0.71
	CDE Raio X	Raio X	0.5	0.55	0.77	0.5	0.5
	CDE TC	TC	0.75	0.75	0.76	0.75	0.75

Fonte: o autor (2023).

Com a finalidade de comparar os resultados obtidos nos experimentos com o estado da arte criamos a Tabela 39 abaixo:

Tabela 39 – Comparação dos resultados com o Estado da Arte onde os datasets foram os mesmos

Modelo	***Dataset***	**Autor**	**AUC**	**Acurácia**	**Precisão**	**Sensibili-dade**	**F1-Score**
CNN	SARS-COV-2 Ct	Castiglione *et al*. (2021)	0,98	0,9999	0,9992	0,9996	-
	SARS-COV-2 Ct	Ferraz & Betini (2023)	0,98	0,89	0,9	0,89	0,89
ViT	HUST-19	Mehboob *et al*. (2022)	-	-	0,997	-	-
	HUST-19	Ferraz & Betini (2023)	1.0	1.0	1.0	1.0	1.0
SwinT	SARS-COV-2 Ct	Peng *et al*., (2022)	0,9991	0,9894	0,9833	0,9895	0,9864
	SARS-COV-2 Ct	Tian *et al*., (2022) (Binária)	0,9985	0,9821	-	0,9907	0,9855
	SARS-COV-2 Ct	Ferraz & Betini (2023)	0,98	0,98	0,98	0,98	0,98

Fonte: o autor (2023).

Ao comparar os resultados obtidos no presente estudo para o modelo CNN com os de Castiglione *et al*. (2021), nota-se que estes últimos obtiveram desempenhos superiores em termos de acurácia, precisão e sensibilidade. Ferraz e Betini (2023), por outro lado, forneceram informações sobre F1-Score, que Castiglione et al. não forneceram.

Em relação ao modelo ViT, comparando os resultados obtidos no presente trabalho com os de Mehboob *et al*. (2022), vemos que os nossos experimentos conseguiram resultados perfeitos em todas as métricas. Mehboob *et al*. (2022) informaram apenas a métrica de precisão, que foi ligeiramente inferior à obtida no presente trabalho.

Por fim, observa-se que os resultados obtidos no presente estudo para o modelo *Swin Transformer* são comparáveis com os obtidos por Peng *et al*. (2022) e Tian *et al*. (2022), embora um pouco inferiores em algumas métricas. No entanto, a diferença é mínima e, dada a alta precisão dos três estudos, os resultados podem ser considerados bastante robustos.

Para concluir, os resultados do presente estudo são promissores e mostram desempenho competitivo quando comparados ao estado da arte. No entanto, é importante destacar que as condições experimentais, as versões dos modelos e os ajustes de hiperparâmetros realizados podem influenciar os resultados. Para uma avaliação mais detalhada, seria ideal ter informações adicionais sobre como cada experimento foi realizado.

4.4.2 Análises Estatísticas e Visuais Sobre os Resultados Obtidos nos Experimentos

4.4.2.1 Análises Estatísticas

Realizamos o Teste de Shapiro-Wilk para avaliar se os pressupostos de normalidade eram satisfeitos para os dados de cada uma das métricas (AUC, acurácia, precisão, sensibilidade e F1-Score). Os resultados estão demonstrados no Quadro 4 e evidenciam que há uma concentração de valores de métricas que não atende aos pressupostos de normalidade para as estratégias *Swin Transformer*. Esses dados serão demonstrados com mais detalhes para cada métrica mais adiante.

Quadro 4 – Métricas que atendem aos pressupostos de normalidade (Teste de Shapiro-Wilk)

Estratégia	Nome da Variável	AUC	Acurácia	Precisão	Sensibilidade	F1-Score
CNN-DS701515-Without_Otimization	C	Não	Sim	Sim	Sim	Sim
CNN-DS701515-Adam	D	Sim	Sim	Sim	Sim	Sim
CNN-DS701515-SGD	E	Não	Sim	Sim	Sim	Sim
CNN-DS602020-Adam	F	Não	Sim	Sim	Sim	Sim
CNN-DS602020-SGD	G	Sim	Sim	Sim	Sim	Sim
SwinT-DS701515-Without_Otimization	H	Não	Não	Não	Não	Não
SwinT-DS701515-Adam	I	Não	Não	Não	Não	Não
SwinT-DS701515-SGD	J	Não	Sim	Não	Não	Não
SwinT-DS602020-Adam	K	Não	Não	Não	Não	Não
SwinT-DS602020-SGD	L	Sim	Sim	Não	Sim	Sim
ViT-DS701515-Without_Otimization	M	Não	Não	Sim	Não	Sim
ViT-DS701515-Adam	N	Sim	Sim	Sim	Sim	Sim
ViT-DS701515-SGD	O	Não	Sim	Não	Sim	Sim
ViT-DS602020-Adam	P	Sim	Sim	Sim	Sim	Sim
ViT-DS602020-SGD	Q	Sim	Sim	Sim	Sim	Sim

Fonte: o autor (2023).

Diante dessas evidências, optamos por realizar testes não paramétricos como o Teste de Wilcoxon e Teste de Friedman para cada uma das métricas. No caso a formulação das hipóteses foi a seguinte:

a) hipótese nula (H0): não há diferença significativa entre os grupos de métricas comparados.

b) hipótese alternativa (H1): pelo menos um grupo de métricas é significativamente diferente dos outros.

Utilizamos a Comparação Par a Par (Durbin-Conover) para comparar os resultados de cada um dos valores das métricas das quinze diferentes estratégias entre si. Optamos por não demonstrar a totalidade desses resultados, por que são 105 valores distintos o que seria difícil de diagramar na estrutura do presente trabalho. Ao invés disso, optamos por demonstrar apenas alguns dos resultados em que o valor-p é menor que o nível de significância pré-definido (0,05), isso porque nesse caso é comum considerar que existe uma diferença estatisticamente significativa entre os grupos comparados, que nesse caso são as diferentes estratégias empregadas para obter os valores de AUC. O Apêndice 1 traz mais informações sobre a Estatística Descritiva para cada uma das cinco métricas.

4.4.2.1.1 Análises Estatísticas da AUC

O Teste de Normalidade dos dados (Shapiro-Wilk), cujos dados constam da Tabela 40, demonstra que a maioria dos dados não seguem uma distribuição normal, portanto é apropriado utilizar testes estatísticos não paramétricos em vez de testes paramétricos.

Tabela 40 – Teste de Normalidade (Shapiro-Wilk) para a AUC

Estratégia	Nome da Variável	Estatística	Valor-p	Resultado
CNN-DS701515-Without_Otimization	C	0.8299	0.0445 (p < 0.05)	Não satisfaz o pressuposto de normalidade
CNN-DS701515-Adam	D	0.8442	0.0644 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-SGD	E	0.8163	0.0313 (p < 0.05)	Não satisfaz o pressuposto de normalidade
CNN-DS602020-Adam	F	0.8045	0.0230 (p < 0.05)	Não satisfaz o pressuposto de normalidade
CNN-DS602020-SGD	G	0.8482	0.0712 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS701515-Without_Otimization	H	0.8022	0.0216 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-Adam	I	0.8280	0.0424 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-SGD	J	0.8322	0.0472 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-Adam	K	0.7841	0.0134 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-SGD	L	0.8380	0.0549 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Without_Otimization	M	0.8039	0.0226 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS701515-Adam	N	0.9241	0.4271 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-SGD	O	0.8240	0.0382 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS602020-Adam	P	0.9460	0.6463 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-SGD	Q	0.9003	0.2538 (p > 0.05)	Satisfaz o pressuposto de normalidade

Fonte: o autor (2023).

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 41 – Teste de Friedman para AUC

χ^2	df	p
32.8	14	0.003

Fonte: o autor (2023).

A Tabela 42 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), analisados na sequência.

Tabela 42 – Comparação Par a Par (Durbin-Conover) para AUC

			Statistic	p
C	–	N	2.6919	0.008
C	-	P	3.9320	< .001
D	-	G	3.1153	0.002
D	-	N	2.6919	0.008
D	-	P	3.9320	< .001
E	-	P	2.2080	0.029
F	-	G	2.4802	0.015
F	-	N	2.0567	0.042
F	-	P	3.2968	0.001
G	-	H	2.2685	0.025
G	-	I	2.7524	0.007
G	-	J	2.6012	0.011
G	-	K	2.8129	0.006
G	-	L	2.1475	0.034
G	-	O	2.3592	0.020
H	-	P	3.0851	0.003
I	-	N	2.3289	0.022
I	-	P	3.5690	< .001
J	-	N	2.1777	0.032
J	-	P	3.4178	< .001
K	-	N	2.3894	0.019
K	-	P	3.6295	< .001
L	-	P	2.9641	0.004
M	-	P	2.6919	0.008
O	-	P	3.1758	0.002

Fonte: o autor (2023).

Agora, pode-se analisar alguns (analisar todos seria dispendioso) dos resultados com as letras substituídas pelas estratégias correspondentes:

a) D – P (CNN-DS701515-Adam x ViT-DS602020-Adam): o valor da estatística é 3.9320 e o valor-p é < .001 (muito baixo). Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias D e P.

b) F – N (CNN-DS602020-Adam x ViT-DS701515-Adam): o valor da estatística é 2.0567 e o valor-p é 0.042. O valor de p, que é 0.042, representa a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso. Em outras palavras, o valor de p é a probabilidade de que a diferença observada entre F e N seja devida ao acaso.

c) H – P (SwinT-DS701515-Without_Optimization x ViT-DS602020-Adam): o valor da estatística é 3.0851 e o valor-p é 0.003. A diferença entre as estratégias H e P é estatisticamente significativa, com um valor de p de 0.003, indicando uma diferença robusta entre essas duas estratégias.

d) I – N (SwinT-DS701515-Adam x ViT-DS701515-Adam): O valor da estatística é 3.5690 e o valor-p é 0.022. A diferença entre as estratégias I e N é estatisticamente significativa, com um valor de p de 0.022, indicando uma diferença moderada entre essas duas estratégias.

e) I – P (SwinT-DS701515-Adam x ViT-DS602020-Adam): O valor da estatística é 3.5690 e o valor-p é < 0.001 (muito baixo). A diferença entre as estratégias I e P é estatisticamente significativa, com um valor de p muito baixo, indicando uma diferença altamente significativa entre essas duas estratégias.

O Teste de Wilcoxon retornou resultados semelhantes conforme a Tabela 43 abaixo:

Tabela 43 – *Paired Samples* Wilcoxon para AUC

			Statistic		p
C	P	Wilcoxon W	44.00		0.013
D	G	Wilcoxon W	45.00		0.009
D	N	Wilcoxon W	34.00	[a]	0.030
D	P	Wilcoxon W	36.00	[a]	0.014
E	G	Wilcoxon W	35.00	[a]	0.021
F	G	Wilcoxon W	45.00		0.009
F	N	Wilcoxon W	33.00	[a]	0.042
F	P	Wilcoxon W	35.00	[a]	0.021
G	J	Wilcoxon W	3.00	[a]	0.041
G	K	Wilcoxon W	4.00		0.032
H	P	Wilcoxon W	35.00	[a]	0.021
I	N	Wilcoxon W	28.00	[b]	0.022
J	N	Wilcoxon W	28.00	[b]	0.022
K	N	Wilcoxon W	34.00	[a]	0.030
K	P	Wilcoxon W	28.00	[b]	0.022
L	P	Wilcoxon W	28.00	[b]	0.022
O	P	Wilcoxon W	28.00	[b]	0.022
Note. H_a $\mu_{\text{Measure 1 - Measure 2}} \neq 0$					
[a] 1 pair(s) of values were tied					
[b] 2 pair(s) of values were tied					

Fonte: o autor (2023).

Convém destacar que a AUC foi a métrica em que o modelo *Swin Transformer* não se destacou da CNN e do ViT, embora isso tenha ocorrido nas demais métricas.

4.4.2.1.2 Análises Estatísticas da Acurácia

O Teste de Normalidade dos dados (Shapiro-Wilk), cujos dados constam da Tabela 44, demonstra que a maioria dos dados não seguem uma distribuição

normal, portanto, é apropriado utilizar testes estatísticos não paramétricos em vez de testes paramétricos.

Tabela 44 – Teste de Normalidade (Shapiro-Wilk) para a Acurácia

Estratégia	Nome da Variável	Estatística	Valor-p	Resultado
CNN-DS701515-Without_Otimization	C	0.8693	0.1210 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-Adam	D	0.9137	0.3430 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-SGD	E	0.9200	0.3924 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-Adam	F	0.8899	0.1992 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-SGD	G	0.8970	0.2353 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS701515-Without_Otimization	H	0.8065	0.0242 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-Adam	I	0.8305	0.0453 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-SGD	J	0.8560	0.0868 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS602020-Adam	K	0.7993	0.0201 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-SGD	L	0.8377	0.0544 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Without_Otimization	M	0.8153	0.0305 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS701515-Adam	N	0.9300	0.4812 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-SGD	O	0.8903	0.2010 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-Adam	P	0.9373	0.5534 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-SGD	Q	0.9055	0.2859 (p > 0.05)	Satisfaz o pressuposto de normalidade

Fonte: o autor (2023).

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 45 – Teste de Friedman para Acurácia

χ^2	df	p
71.8	14	< .001

Fonte: o autor (2023).

A Tabela 46 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), que serão analisados na sequência.

Tabela 46 – Comparação Para a Par (Durbin-Conover) para Acurácia

			Statistic	p
C	-	H	3.5990	< .001
D	-	E	2.1284	0.035
D	-	G	2.7863	0.006
D	-	H	2.3993	0.018
D	-	I	3.4829	< .001

D	-	O	2.6702	0.009
E	-	H	4.5277	< .001
E	-	L	4.7986	< .001
F	-	H	3.6376	< .001
F	-	I	4.7212	< .001
F	-	J	4.9147	< .001
F	-	K	4.7986	< .001
F	-	L	3.9085	< .001
F	-	Q	3.0959	0.002
G	-	H	5.1856	< .001
G	-	Q	4.6438	< .001
H	-	M	3.9085	< .001
H	-	N	3.0959	0.002
H	-	P	3.2894	0.001
I	-	M	4.9921	< .001
I	-	N	4.1794	< .001
J	-	M	5.1856	< .001
J	-	N	4.3729	< .001
J	-	P	4.5664	< .001
K	-	M	5.0695	< .001
K	-	N	4.2568	< .001
K	-	P	4.4503	< .001
L	-	M	4.1794	< .001
M	-	O	4.1794	< .001
M	-	Q	3.3668	0.001
P	-	Q	2.7476	0.007

Fonte: o autor (2023).

Agora, pode-se analisar alguns dos resultados com as letras substituídas pelas estratégias correspondentes:

a) C – H (CNN-DS701515-Without_Otimization x SwinT-DS701515-Without_Otimization): o valor da estatística é 3.5990 e o valor-p é < .001 (muito baixo). Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias C e H.

b) D – G (CNN-DS701515-Adam x CNN-DS602020-SGD): o valor da estatística é 2.7863 e o valor-p é 0.006. Esse valor de p representa que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre D e G.

c) H – M (SwinT-DS701515-Without_Optimization x ViT-DS701515-Without_Otimization): o valor da estatística é 3.9085 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito

fortes de que existe uma diferença altamente significativa entre as estratégias H e M.

d) K – N (SwinT-DS602020-Adam x ViT-DS701515-Adam): o valor da estatística é 4.2568 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias K e N.

e) L – M (SwinT-DS602020-SGDxViT-DS701515-Without_Otimization): o valor da estatística é 4.1794 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias

O Teste de Wilcoxon retornou resultados semelhantes conforme a Tabela 47 abaixo:

Tabela 47 – Teste de Wilcoxon para a Acurácia

			Statistic		p
C	H	Wilcoxon W	1.50		0.015
H	M	Wilcoxon W	45.00		0.009
D	I	Wilcoxon W	0.00		0.009
D	E	Wilcoxon W	36.00	[a]	0.013
D	G	Wilcoxon W	45.00		0.009
H	N	Wilcoxon W	42.00		0.020
H	P	Wilcoxon W	35.00	[a]	0.021
I	M	Wilcoxon W	45.00		0.009
I	N	Wilcoxon W	27.00	[b]	0.035
J	P	Wilcoxon W	36.00	[a]	0.014
J	M	Wilcoxon W	45.00		0.009
K	N	Wilcoxon W	44.00		0.013
K	N	Wilcoxon W	35.00	[a]	0.021
K	P	Wilcoxon W	36.00	[a]	0.014
P	Q	Wilcoxon W	0.00	[a]	0.014
F	K	Wilcoxon W	0.00		0.004
F	Q	Wilcoxon W	3.00		0.020

Note. H_a $\mu_{\text{Measure 1 - Measure 2}} \neq 0$

[a] 1 pair(s) of values were tied

[b] 2 pair(s) of values were tied

Fonte: o autor (2023).

4.4.2.1.3 Análises Estatísticas da Precisão

O Teste de Normalidade dos dados (Shapiro-Wilk), cujos dados constam da Tabela 48, demonstra que a maioria dos dados não seguem uma distribuição

normal, portanto é apropriado utilizar testes estatísticos não paramétricos em vez de testes paramétricos.

Tabela 48 – Teste de Normalidade (Shapiro-Wilk) para a Precisão

Estratégia	Nome da Variável	Estatística	Valor-p	Resultado
CNN-DS701515-Without_Otimization	C	0.8985	0.2436 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-Adam	D	0.9339	0.5192 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-SGD	E	0.8763	0.1436 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-Adam	F	0.8526	0.0796 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-SGD	G	0.9280	0.4624 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS701515-Without_Otimization	H	0.7905	0.0159 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-Adam	I	0.7857	0.0140 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-SGD	J	0.7293	0.0031 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-Adam	K	0.8133	0.0289 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-SGD	L	0.8034	0.0223 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS701515-Without_Otimization	M	0.8393	0.0567 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Adam	N	0.8809	0.1606 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-SGD	O	0.7921	0.0166 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS602020-Adam	P	0.9589	0.7868 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-SGD	Q	0.8667	0.1134 (p > 0.05)	Satisfaz o pressuposto de normalidade

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 49 – Teste de Friedman para Precisão

X^2	df	p
57.4	14	< .001

Fonte: o autor (2023).

A Tabela 50 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), que serão analisados na sequência.

Tabela 50 – Comparação Par a Par (Durbin-Conover)

			Statistic	P
C	-	H	2.3283	0.022
C	-	I	3.2185	0.002
D	-	G	2.3968	0.018
D	-	I	2.4652	0.015
E	-	H	3.1500	0.002
E	-	I	4.0402	< .001

E	-	Q	2.2940	0.024
F	-	H	2.5337	0.013
F	-	I	3.4239	< .001
G	-	H	3.9718	< .001
G	-	I	4.8620	< .001
G	-	J	4.8277	< .001
G	-	K	5.6153	< .001
G	-	L	4.9647	< .001
G	-	O	3.6294	< .001
G	-	Q	3.1158	0.002
H	-	M	2.8419	0.005
H	-	N	3.0473	0.003
H	-	P	2.7391	0.007
I	-	M	3.7321	< .001
I	-	N	3.9375	< .001
I	-	P	3.6294	< .001
J	-	M	3.6978	< .001
J	-	N	3.9033	< .001
J	-	P	3.5951	< .001
K	-	M	4.4854	< .001
K	-	N	4.6908	< .001
K	-	O	1.9859	0.049
K	-	P	4.3826	< .001
K	-	Q	2.4995	0.014
L	-	M	3.8348	< .001
L	-	N	4.0402	< .001
L	-	P	3.7321	< .001
M	-	O	2.4995	0.014
N	-	O	2.7049	0.008
N	-	Q	2.1913	0.030
O	-	P	2.3968	0.018

Fonte: o autor (2023).

Agora, pode-se analisar alguns dos resultados com as letras substituídas pelas estratégias correspondentes:

a) C – H (CNN-DS701515-Without_Otimization x SwinT-DS701515-Without_Otimization): o valor da estatística é 2.3283 e o valor-p é 0.022. Esse valor de p indica que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre as estratégias C e H.

b) D – G (CNN-DS701515-Adam x CNN-DS602020-SGD): o valor da estatística é 2.3968 e o valor-p é 0.018. Esse valor de p representa que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre D e G.

c) H – M (SwinT-DS701515-Without_Optimization x ViT-DS701515-Without_Otimization): o valor da estatística é 2.8419 e o valor-p é 0.005. Esse valor de p indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas da existência de diferença estatisticamente significativas entre as estratégias H e M.

d) K – N (SwinT-DS602020-Adam x ViT-DS701515-Adam): o valor da estatística é 4.6908 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias K e N.

e) L – M (SwinT-DS602020-SGDxViT-DS701515-Without_Otimization): o valor da estatística é 3.8348 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias.

O Teste de Wilcoxon retornou resultados semelhantes conforme a Tabela 51 abaixo:

Tabela 51 – Teste de Wilcoxon para a Precisão

			Statistic		**p**
H	M	Wilcoxon W	35.00	[a]	0.020
H	N	Wilcoxon W	36.00		0.123
D	I	Wilcoxon W	4.00		0.033
D	E	Wilcoxon W	39.00		0.057
D	G	Wilcoxon W	45.00		0.009
I	M	Wilcoxon W	43.00		0.012
I	N	Wilcoxon W	35.00	[a]	0.021
K	M	Wilcoxon W	44.00		0.013
K	N	Wilcoxon W	36.00	[a]	0.014
K	Q	Wilcoxon W	36.00	[a]	0.014
K	P	Wilcoxon W	36.00	[a]	0.014
L	M	Wilcoxon W	45.00		0.009
L	N	Wilcoxon W	36.00	[a]	0.014
C	I	Wilcoxon W	1.50		0.015
E	H	Wilcoxon W	6.00		0.055
E	I	Wilcoxon W	3.50		0.028
E	Q	Wilcoxon W	5.50		0.050
Note. **H_a $\mu_{\text{Measure 1 - Measure 2}} \neq 0$**					
[a] 1 *pair(s) of values were tied*					

Fonte: o autor (2023).

4.4.2.1.4 Análises Estatísticas da Sensibilidade/Recall

O Teste de Normalidade dos dados (Shapiro-Wilk), cujos dados constam da Tabela 52, demonstra que a maioria dos dados não seguem uma distribuição

normal, portanto é apropriado utilizar testes estatísticos não paramétricos em vez de testes paramétricos.

Tabela 52 – Teste de Normalidade (Shapiro-Wil) para a Sensibilidade

Estratégia	Nome da Variável	Estatística	Valor-p	Resultado
CNN-DS701515-Without_Otimization	C	0.8693	0.1210 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-Adam	D	0.9168	0.3666 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-SGD	E	0.9097	0.3140 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-Adam	F	0.8910	0.2044 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-SGD	G	0.9086	0.3061 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS701515-Without_Otimization	H	0.8065	0.0242 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-Adam	I	0.8280	0.0424 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-SGD	J	0.8322	0.0472 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-Adam	K	0.7841	0.0134 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-SGD	L	0.8380	0.0549 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Without_Otimization	M	0.8153	0.0305 (p < 0.05)	Não satisfaz o pressuposto de normalidade
ViT-DS701515-Adam	N	0.9056	0.2860 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-SGD	O	0.8983	0.2423 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-Adam	P	0.9118	0.3286 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-SGD	Q	0.8974	0.2371 (p > 0.05)	Satisfaz o pressuposto de normalidade

Fonte: o autor (2023).

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 53 – Teste de Friedman para Sensibilidade

χ^2	df	p
67.4	14	< .001

Fonte: o autor (2023).

A Tabela 54 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), que serão analisados na sequência.

Tabela 54 – Comparação Par a Par (Durbin-Conover) para Sensibilidade

			Statistic	p
C	-	H	3.8316	< .001
C	-	I	3.9060	< .001
C	-	J	4.1292	< .001
D	-	G	2.3064	0.023
D	-	H	3.6456	< .001
D	-	I	3.7200	< .001
E	-	H	4.9475	< .001

E	-	I	5.0219	< .001
F	-	H	4.0548	< .001
F	-	I	4.1292	< .001
G	-	H	5.9519	< .001
G	-	O	5.0591	< .001
H	-	M	3.7944	< .001
H	-	N	3.7944	< .001
H	-	P	3.9432	< .001
I	-	M	3.8688	< .001
I	-	N	3.8688	< .001
I	-	P	4.0176	< .001
J	-	M	4.0920	< .001
J	-	N	4.0920	< .001
J	-	P	4.2408	< .001
K	-	N	3.9804	< .001
K	-	P	4.1292	< .001
L	-	M	3.3108	0.001
L	-	N	3.3108	0.001
L	-	P	3.4596	< .001
M	-	O	2.9016	0.004
M	-	Q	2.3064	0.023
N	-	O	2.9016	0.004
N	-	Q	2.3064	0.023
O	-	P	3.0504	0.003
P	-	Q	2.4552	0.016

Fonte: o autor (2023).

Agora, pode-se analisar alguns dos resultados com as letras substituídas pelas estratégias correspondentes:

a) C – H (CNN-DS701515-Without_Otimization x SwinT-DS701515-Without_Otimization): o valor da estatística é 3.8316 e o valor-p é < .001 (muito baixo). Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias C e H.

b) D – G (CNN-DS701515-Adam x CNN-DS602020-SGD): oO valor da estatística é 2.3064 e o valor-p é 0.023. Esse valor de p representa que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre D e G.

c) H – M (SwinT-DS701515-Without_Optimization x ViT-DS701515-Without_Otimization): o valor da estatística é 3.7944 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito

fortes de que existe uma diferença altamente significativa entre as estratégias H e M.

d) K – N (SwinT-DS602020-Adam x ViT-DS701515-Adam): o valor da estatística é 3.9804 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias K e N.

e) L – M (SwinT-DS602020-SGDxViT-DS701515-Without_Otimization): o valor da estatística é 3.3108 e o valor-p é 0.001. Esse valor de p "0.001" indica que existe uma diferença significativa entre as estratégias L e M.

O Teste de Wilcoxon retornou resultados semelhantes conforme a Tabela 55 abaixo:

Tabela 55 – Teste de Wilcoxon para a Sensibilidade

			Statistic		**p**
C	H	Wilcoxon W	1.50		0.015
C	I	Wilcoxon W	1.50		0.015
C	J	Wilcoxon W	4.00		0.032
C	K	Wilcoxon W	4.00		0.033
C	L	Wilcoxon W	3.00		0.024
D	H	Wilcoxon W	1.00	[a]	0.021
D	I	Wilcoxon W	0.00		0.009
D	J	Wilcoxon W	0.00	[a]	0.014
D	K	Wilcoxon W	0.00	[a]	0.014
D	L	Wilcoxon W	0.00	[a]	0.014
I	M	Wilcoxon W	44.00		0.013
K		Wilcoxon W	42.00		0.020
K	N	Wilcoxon W	35.00	[a]	0.021
K	P	Wilcoxon W	28.00	[b]	0.022
L	M	Wilcoxon W	43.00		0.017
L	N	Wilcoxon W	35.00	[a]	0.021
F	I	Wilcoxon W	2.00		0.018

Note. H_a μ Measure 1 - Measure 2 ≠ 0

[a] **1** ***pair(s) of values were tied***

[b] **2** ***pair(s) of values were tied***

Fonte: o autor (2023).

4.4.2.1.5 Análises Estatísticas da F1-Score

O Teste de Normalidade dos dados (Shapiro-Wilk), cujos dados constam da Tabela 56, demonstra que a maioria dos dados não seguem uma distribuição normal, portanto é apropriado utilizar testes estatísticos não paramétricos em vez de testes paramétricos.

Tabela 56 – Teste de Normalidade (Shapiro-Wilk) para a F1-Score

Estratégia	Nome da Variável	Estatística	Valor-p	Resultado
CNN-DS701515-Without_Otimization	C	0.8796	0.1554 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-Adam	D	0.9055	0.2859 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS701515-SGD	E	0.8985	0.2433 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-Adam	F	0.8765	0.1441 (p > 0.05)	Satisfaz o pressuposto de normalidade
CNN-DS602020-SGD	G	0.9340	0.5208 (p > 0.05)	Satisfaz o pressuposto de normalidade
SwinT-DS701515-Without_Otimization	H	0.7861	0.0141 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-Adam	I	0.8004	0.0207 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS701515-SGD	J	0.7706	0.0094 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-Adam	K	0.7772	0.0112 (p < 0.05)	Não satisfaz o pressuposto de normalidade
SwinT-DS602020-SGD	L	0.8380	0.0549 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Without_Otimization	M	0.8346	0.0503 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-Adam	N	0.9036	0.2739 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS701515-SGD	O	0.8611	0.0987 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-Adam	P	0.9371	0.5513 (p > 0.05)	Satisfaz o pressuposto de normalidade
ViT-DS602020-SGD	Q	0.9050	0.2823 (p > 0.05)	Satisfaz o pressuposto de normalidade

Fonte: o autor (2023).

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 57 – Teste de Friedman para F1-Score

χ^2	df	p
58.4	14	< .001

Fonte: o autor (2023).

A Tabela 58 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), que serão analisados na sequência.

Tabela 58 – Comparação Par a Par (Durbin-Conover) para F1-Score

			Statistic	p
C	-	H	3.5844	< .001
C	-	I	3.6533	< .001
C	-	J	4.0324	< .001
D	-	H	2.9295	0.004
D	-	I	2.9985	0.003
D	-	J	3.3776	0.001
E	-	H	4.2737	< .001
E	-	I	4.3426	< .001
E	-	J	4.7217	< .001
F	-	H	3.1019	0.002
F	-	I	3.1708	0.002

F	-	J	3.5499	< .001
F	-	K	4.0669	< .001
G	-	H	3.9980	< .001
G	-	I	4.0669	< .001
G	-	J	4.4460	< .001
G	-	K	4.9630	< .001
H	-	M	3.2397	0.002
H	-	N	2.7572	0.007
H	-	P	2.3781	0.019
I	-	M	3.3087	0.001
I	-	N	2.8261	0.006
I	-	P	2.4470	0.016
J	-	M	3.6878	< .001
J	-	N	3.2053	0.002
J	-	P	2.8261	0.006
K	-	M	4.2048	< .001
K	-	N	3.7222	< .001
K	-	P	3.3431	0.001
L	-	M	3.8256	< .001
L	-	N	3.3431	0.001
L	-	P	2.9640	0.004
M	-	O	2.5160	0.013
M	-	Q	2.8261	0.006
N	-	O	2.0334	0.044
N	-	Q	2.3436	0.021
P	-	Q	1.9645	0.052

Fonte: o autor (2023).

Agora, pode-se analisar alguns dos resultados com as letras substituídas pelas estratégias correspondentes:

a) C – H (CNN-DS701515-Without_Otimization x SwinT-DS701515-Without_Otimization): o valor da estatística é 3.5844 e o valor-p é < .001 (muito baixo). Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias C e H.

b) D – H (CNN-DS701515-Adam x CNN-DS701515-Without_Otimization): o valor da estatística é 2.9295 e o valor-p é 0.004. Esse valor de p representa que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre D e H.

c) H – M (SwinT-DS701515-Without_Optimization x ViT-DS701515-Without_Otimization): o valor da estatística é 3.2397 e o valor-p é 0.002. Esse valor de p representa que há evidências estatísticas para concluir que existe uma diferença estatisticamente significativa entre as estratégias H e M.

d) K – N (SwinT-DS602020-Adam x ViT-DS701515-Adam): o valor da estatística é 3.7222 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias K e N.

e) L – M (SwinT-DS602020-SGDxViT-DS701515-Without_Otimization): o valor da estatística é 3.8256 e o valor-p é < .001. Esse valor de p "< 0.001" indica que a probabilidade de obter uma diferença tão grande ou maior do que a observada, devido ao acaso, é extremamente baixa. Em outras palavras, há evidências estatísticas muito fortes de que existe uma diferença altamente significativa entre as estratégias.

O Teste de Wilcoxon retornou resultados semelhantes conforme a Tabela 59 abaixo:

Tabela 59 – Teste de Wilcoxon para a Sensibilidade

			Statistic		p
C	H	Wilcoxon W	2.00		0.017
C	I	Wilcoxon W	1.00		0.013
C	J	Wilcoxon W	4.00		0.032
D	H	Wilcoxon W	1.00	[a]	0.035
D	I	Wilcoxon W	1.50		0.015
D	J	Wilcoxon W	2.00		0.012
E	H	Wilcoxon W	3.50		0.028
E	I	Wilcoxon W	0.00	[b]	0.014
E	J	Wilcoxon W	1.50		0.015
F	I	Wilcoxon W	2.00	[a]	0.051
F	K	Wilcoxon W	0.00		0.009
G		Wilcoxon W	0.00	[b]	0.014
H	M	Wilcoxon W	41.00		0.033
K	N	Wilcoxon W	35.00	[b]	0.021
K	M	Wilcoxon W	44.00		0.008
K	P	Wilcoxon W	34.50	[b]	0.025
L		Wilcoxon W	35.00	[b]	0.021

Note. H_a μ Measure 1 - Measure 2 ≠ 0

[a] ***2 pair(s) of values were tied***

[b] ***1 pair(s) of values were tied***

Fonte: o autor (2023).

Por fim, cabe destacar que o foram realizados os testes estatísticos sobre a mediana dos valores das métricas descritas na Tabela 60.

Tabela 60 – Medianas de todas as estratégias agrupadas por modelo e métricas

Modelos/Nome Variável			
Métrica	**CNN/B**	**SwinT/C**	**ViT/D**
AUC	0.85	0.85	0.81
Acurácia	0.85	0.94	0.84
Precisão	0.85	0.94	0.84
Sensibilidade	0.85	0.94	0.78
F1-Score	0.85	0.94	0.84

Fonte: o autor (2023).

O Teste de Friedman executado sugere que existe evidência estatística para rejeitar a hipótese nula de que não há diferença significativa entre as métricas comparadas. Em outras palavras, os grupos são diferentes em termos estatisticamente significativos.

Tabela 61 – Teste de Friedman os três modelos

χ^2	df	p
9.58	2	0.008

Fonte: o autor (2023).

A Tabela 62 demonstra os dados obtidos na Comparação Par a Par (Durbin-Conover), que serão analisados na sequência.

Tabela 62 – Comparação Par a Par (Durbin-Conover) para os três modelos

			Statistic	p
B	-	C	5.66	< .001
B	-	D	7.78	< .001
C	-	D	13.44	< .001

Fonte: o autor (2023).

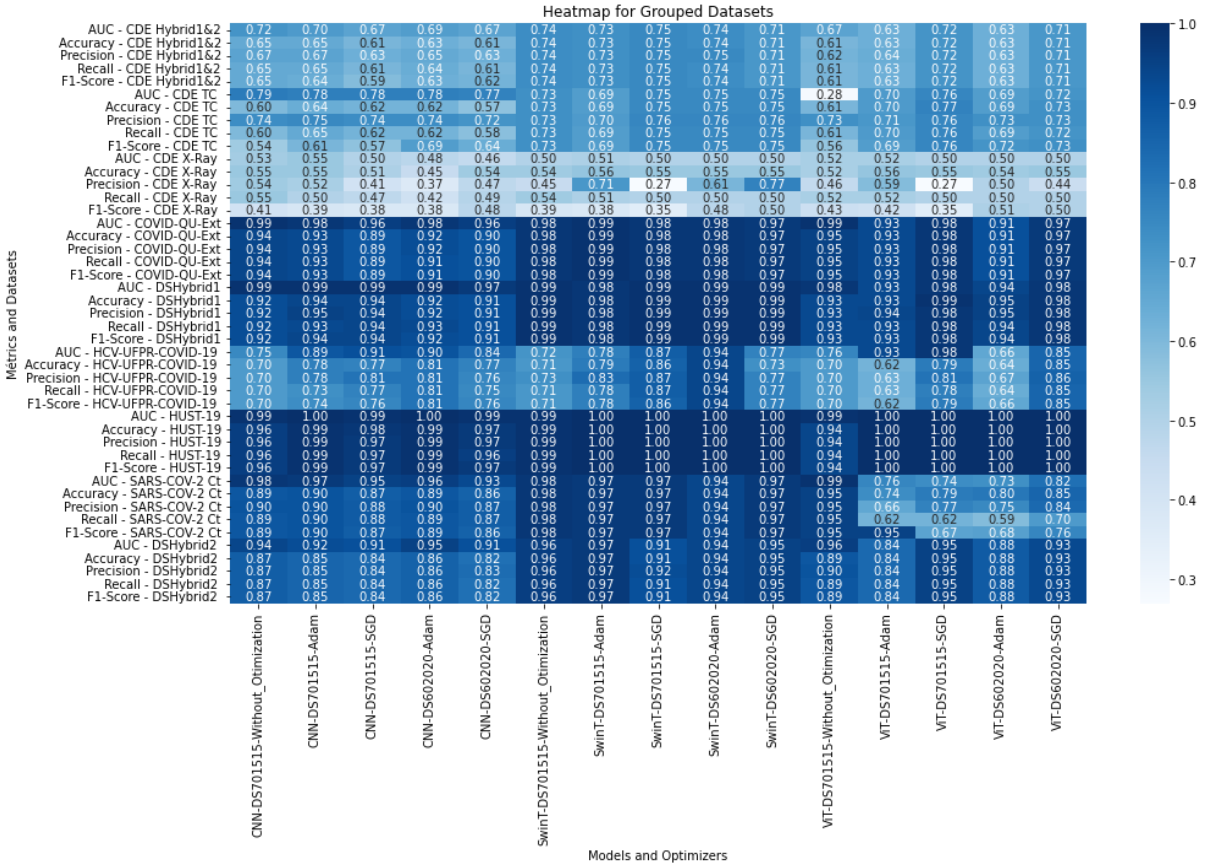
Todas as comparações em pares (B-C, B-D e C-D) mostram valores de valor-p muito pequenos, todos indicados como <.001. Isso significa que, em todas as comparações, as diferenças observadas entre as estratégias são altamente improváveis de ocorrerem ao acaso. Portanto, com base nessas análises, pode-se afirmar que há evidências estatísticas muito fortes para concluir que existem diferenças significativas entre os modelos CNN, *Swin Transformer* e *Vision Transformer*.

Essas diferenças não são simplesmente devidas ao acaso, mas são estatisticamente altamente significativas.

Esses resultados indicam que pelo menos uma das comparações em pares entre os modelos CNN, *Swin Transformer* e *Vision Transformer* é estatisticamente significativa, embora não forneçam informações específicas sobre quais comparações individuais são significantes.

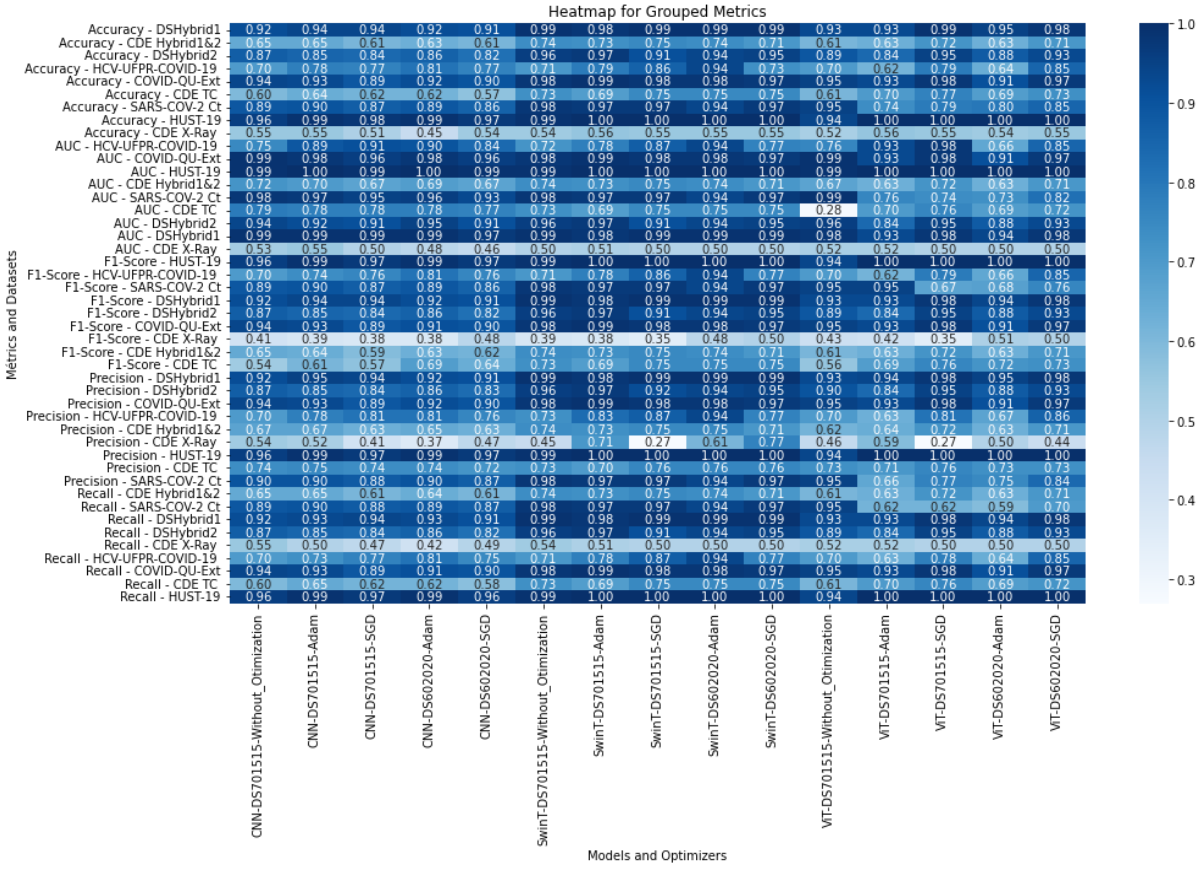
4.4.2.2 Análises Exploratórias de Dados (AED)

Figura 12 – *Heatmap* dos valores agrupados por *datasets* e estratégias



Fonte: o autor (2023).

Figura 13 – *Heatmap* dos valores agrupados por métricas e estratégias

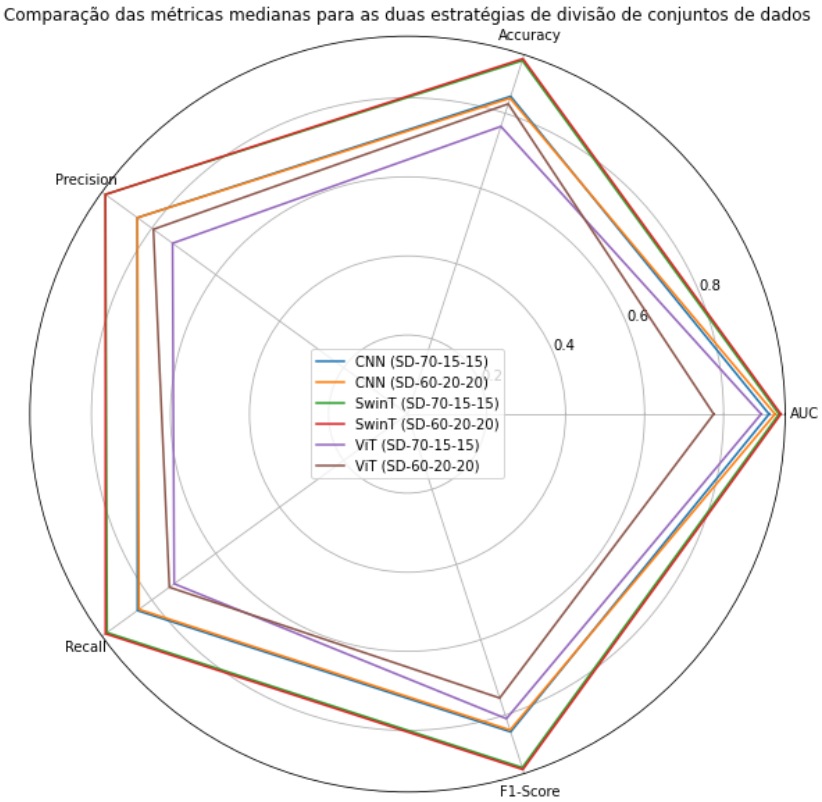


Fonte: o autor (2023).

Analisando as Figuras 12 e 13 podemos verificar, facilmente, que os valores das métricas obtidas pelos modelos *Swin Transformer* são conjuntamente mais elevados na maioria das estratégias de execução. De forma rápida é possível verificar que os resultados obtidos pelo *Swin Transformer* são mais elevados que os do ViT e do CNN em pelo menos sete datasets: HUST-19, SARS-COV2-Ct, DSHybrid1, DSHybrid2, COVID-QU-Ext, CDE Hybrid1&2 e CDE TC.

O Gráfico 2 ajuda a identificar como os modelos se comportaram nos diferentes *split datasets*, no caso 70/15/15 e 60/20/20.

Gráfico 2 – Radar para as medianas nos *split datasets* 70/15/15 e 60/20/20

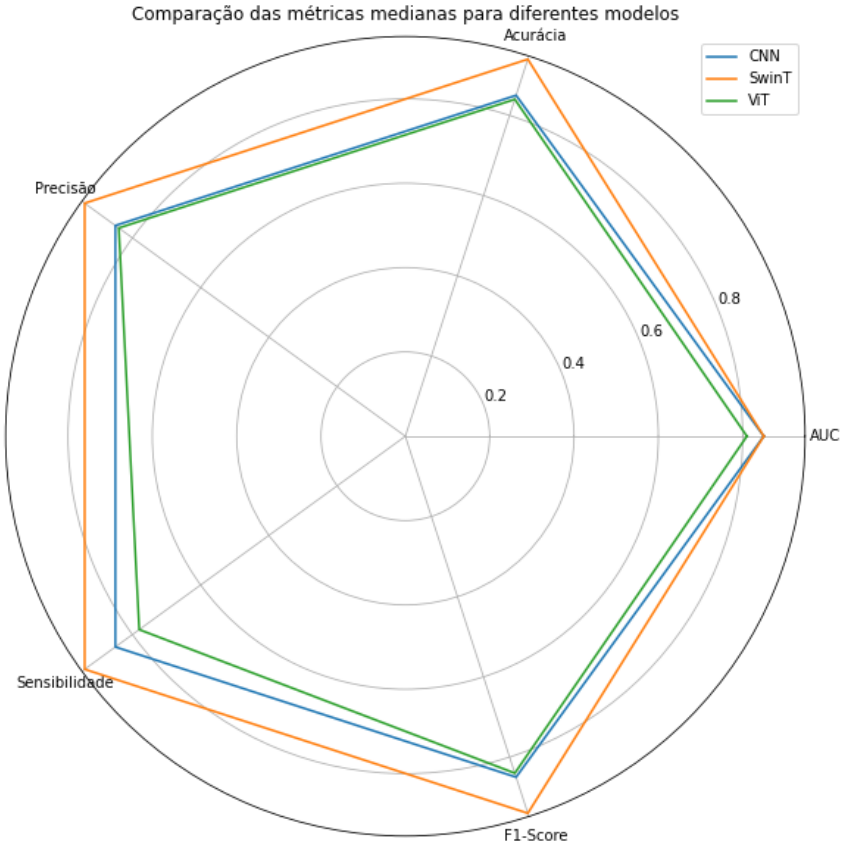


Fonte: o autor (2023).

Analisando o gráfico podemos verificar que o modelo que mais sofreu variação nas métricas nas diferentes estratégias de *split dataset* foi o ViT. Outro ponto notável é que a AUC é a métrica onde os modelos obtiveram resultados mais assemelhados, com exceção do ViT com o *split dataset* 60/20/20. Outro ponto notável é que o *Swin Transformer* obteve os maiores valores absolutos em todas as métricas, embora na AUC essa vantagem tenha sido mínima.

O Gráfico 3 traz uma nova visão onde os dados estão consolidados em uma única mediana para cada métrica e modelo. Nesse caso, ainda observamos uma liderança do modelo *Swin Transformer*, em todas as métricas, com exceção da AUC em que ele empata com o CNN. Nas demais métricas o CNN e ViT disputam o segundo lugar com leve vantagem para o CNN.

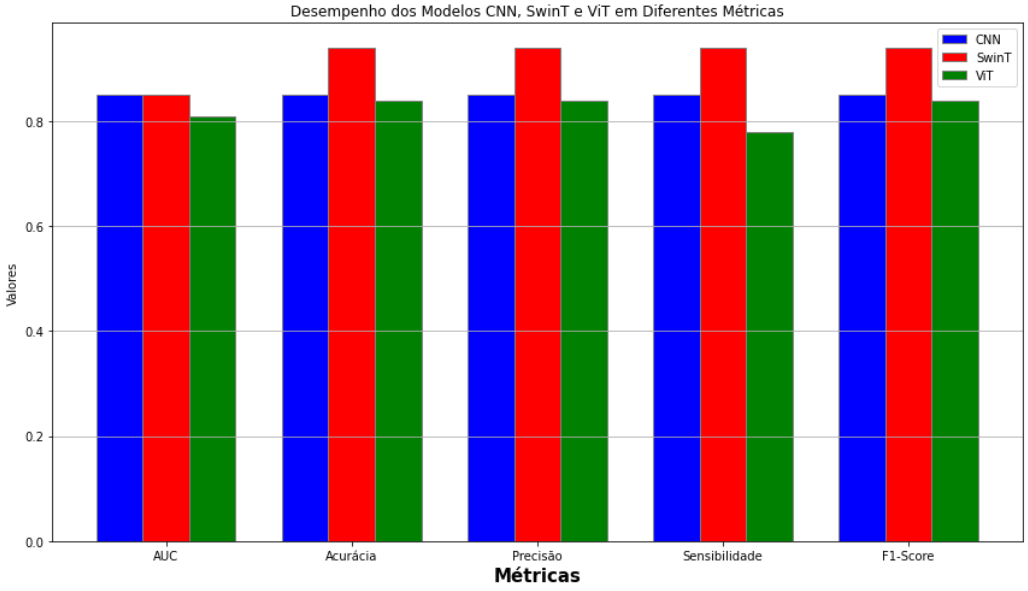
Gráfico 3 – Radar para as medianas dos modelos



Fonte: o autor (2023).

O Gráfico 4 traz os valores das medianas obtidas nas cinco métricas pelos três modelos.

Gráfico 4 – Medianas das cinco métricas obtidas pelos modelos



Fonte: o autor (2023).

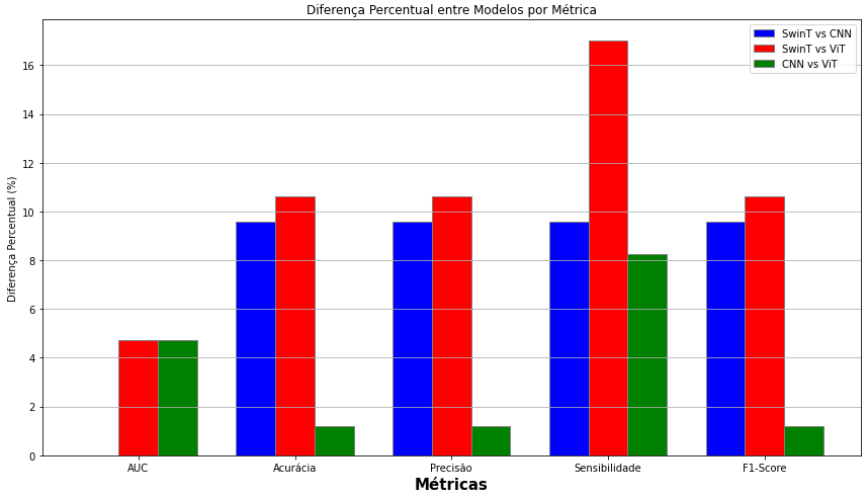
Realizamos uma comparação de diferenças entre os as medianas das métricas obtidas por cada um dos modelos. Para facilitar a análise e visualização compilamos os dados na Tabela 63 e criamos quatro gráficos (5, 6, 7 e 8).

Tabela 63 – Diferença percentual das métricas entre os modelos

Métrica	Diferença % (SwinT - CNN)	Diferença % (SwinT - ViT)	Diferença % (CNN - ViT)
AUC	0.0%	4.94%	4.94%
Acurácia	10.59%	11.9%	1.19%
Precisão	10.59%	11.9%	1.19%
Sensibilidade	10.59%	20.51%	8.97%
F1-Score	10.59%	11.9%	1.19%

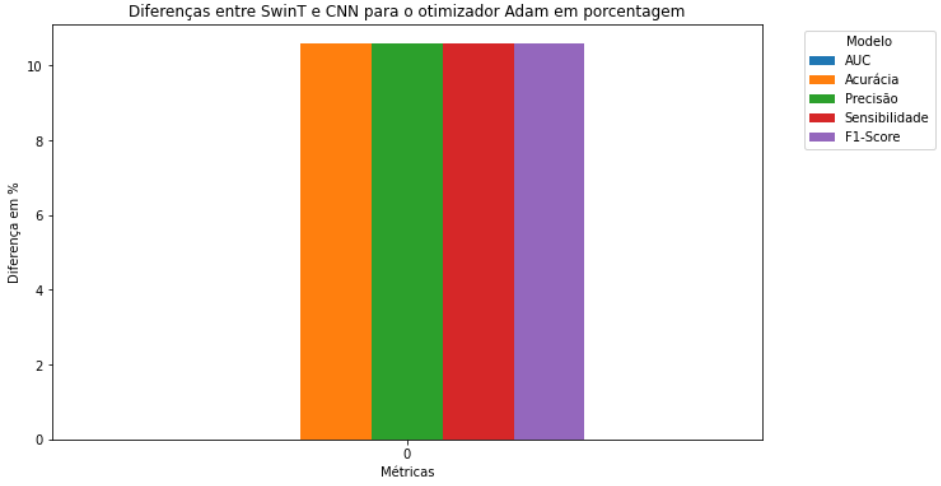
Fonte: o autor (2023).

Gráfico 5 – Diferença (%) entre todos os modelos por métrica



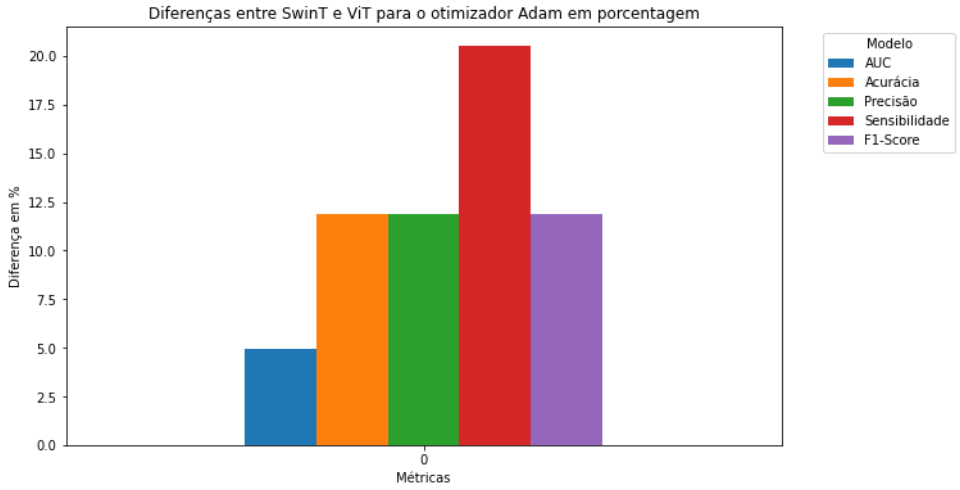
Fonte: o autor (2023).

Gráfico 6 – Diferença (%) entre *SwinT* e CNN



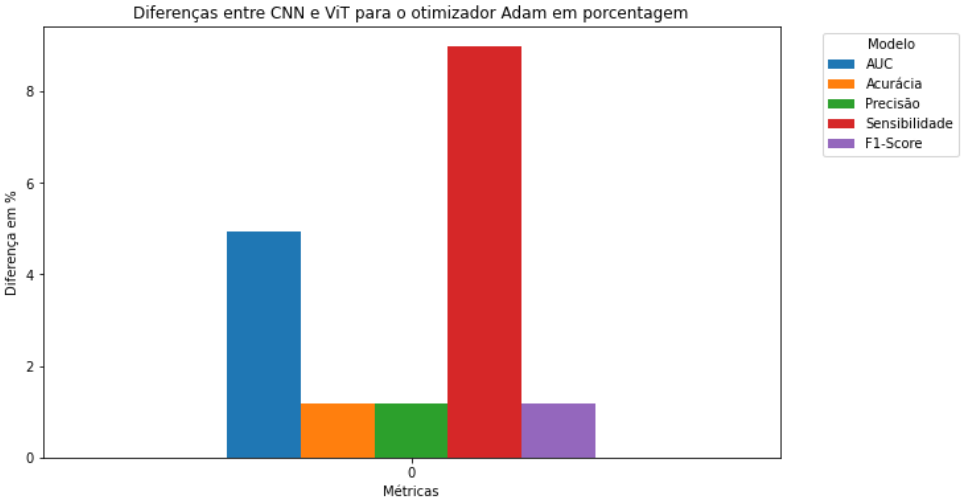
Fonte: o autor (2023).

Gráfico 7 – Diferença (%) entre *SwinT* e ViT



Fonte: o autor (2023).

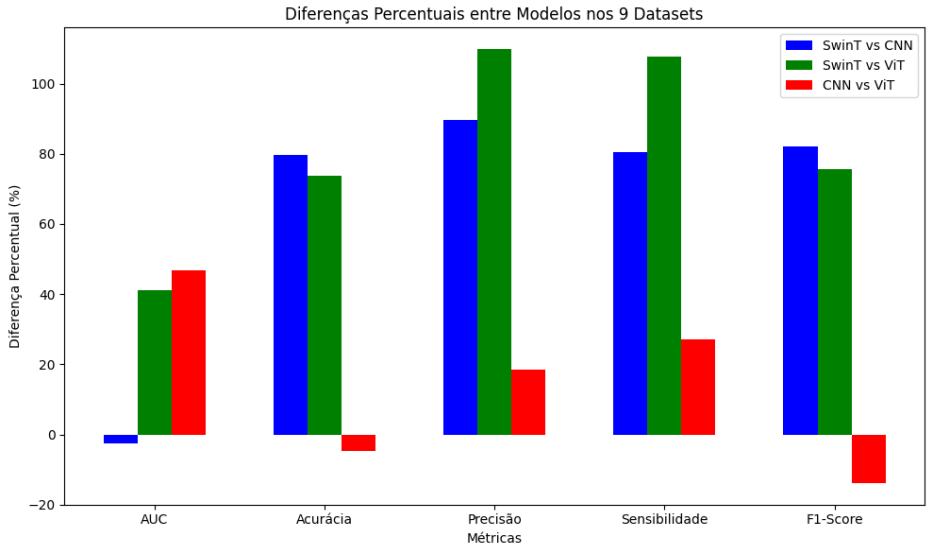
Gráfico 8 – Diferença (%) entre CNN e ViT



Fonte: o autor (2023).

O Gráfico 9 demonstra o consolidado das diferenças percentuais das cinco métricas nos nove *datasets.* A análise desses resultados de desempenho dos modelos revela informações essenciais. Começando pela métrica AUC, observamos que o modelo *SwinT* apresenta uma ligeira desvantagem em relação ao modelo CNN, com uma diferença de -2.55%, indicando que o CNN é melhor nessa métrica. No entanto, o *SwinT* supera significativamente o ViT com uma diferença de 41.19%, e o CNN supera o ViT com uma diferença ainda maior de 46.81%.

Gráfico 9 – Diferença (%) entre modelos nos nove *datasets*



Fonte: o autor (2023).

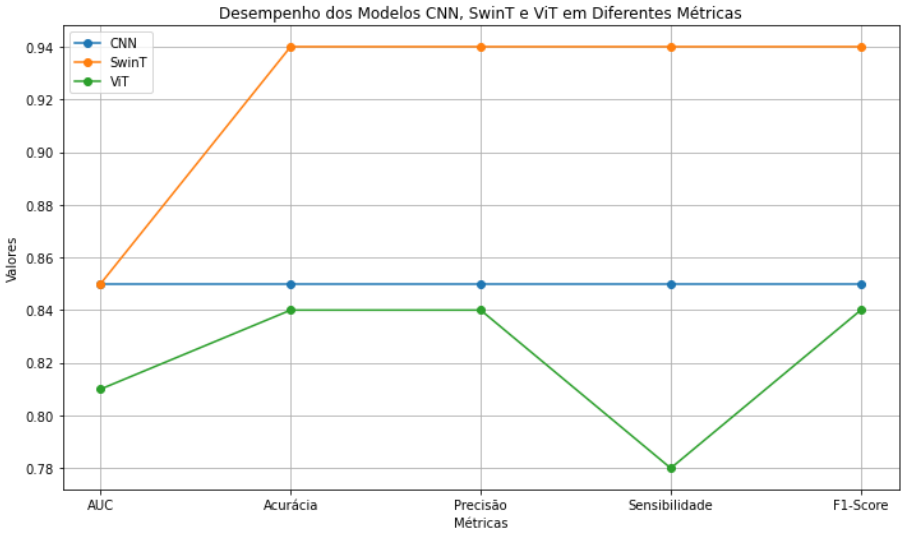
Na métrica de acurácia, o *SwinT* lidera com 79.79% em comparação com o CNN, enquanto o ViT apresenta a acurácia mais baixa, com -4.81% em relação ao CNN. Quando olhamos para a precisão, o *SwinT* também supera o CNN, com uma precisão superior em 89.71%. Quando comparado com o ViT, o *SwinT* lidera com uma precisão de 109.81% superior. O CNN supera o ViT com uma precisão 18.56% maior.

A métrica de sensibilidade (*recall*) mostra que o *SwinT* tem uma sensibilidade de 80.62% superior em comparação com o CNN. Quando comparado com o ViT, o *SwinT* lidera com uma precisão de 107.80% superior. Mais uma vez, o CNN ocupa a última posição, com uma sensibilidade 27.08% maior do que o ViT. Finalmente, no F1-Score, o *SwinT* lidera com 82.19% contra o CNN, também fica na frente do ViT, que possui um F1-Score de 75.60% inferior ao SwinT. O CNN tem o desempenho mais baixo nessa métrica, com um F1-Score 13.83% menor em relação ao ViT.

Observando os resultados, percebe-se que o *Swin Tranformer* apresentou um desempenho global 2,55% menor que a CNN. Grande parte dessa diferença decorre do desempenho desfavorável do *Swin Transformer* no *dataset* HCV-UFPR-COVID-19, onde estava 12,36% atrás da CNN. Esse resultado comparativo representa o ponto mais baixo do *Swin Transformer* ao longo de todo o estudo.

Abaixo temos o Gráfico 10 que demonstra as medianas das métricas consolidada para cada modelo:

Gráfico 10 – Medianas das métricas dos modelos



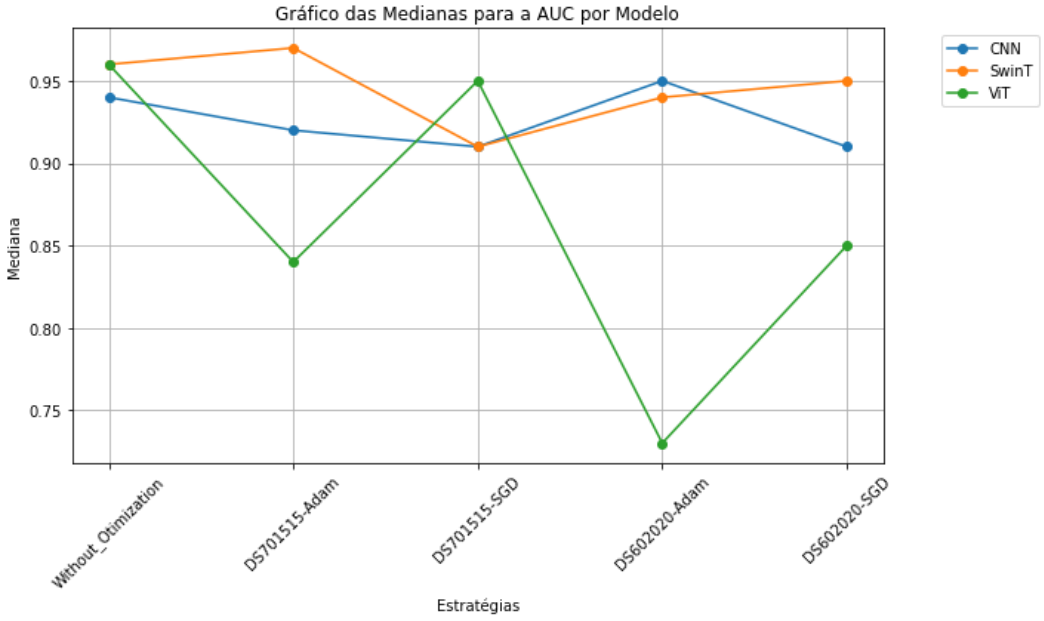
Fonte: o autor (2023).

De forma geral, ao analisarmos os dados apresentados é possível verificar que os três modelos têm bom desempenho, porém o *Swin Transformer* se destaca em quatro das cinco métricas avaliadas.

4.4.2.2.1 AED da AUC

O Gráfico 11 demonstra que a AUC teve diferentes resultados nas diferentes estratégias e modelos. Pode-se verificar que os modelos CNN e *Swin Transformer* obtiveram valores que oscilaram menos que o modelo ViT quando analisadas as cinco estratégias.

Gráfico 11 – Medianas da AUC para as quinze diferentes estratégias

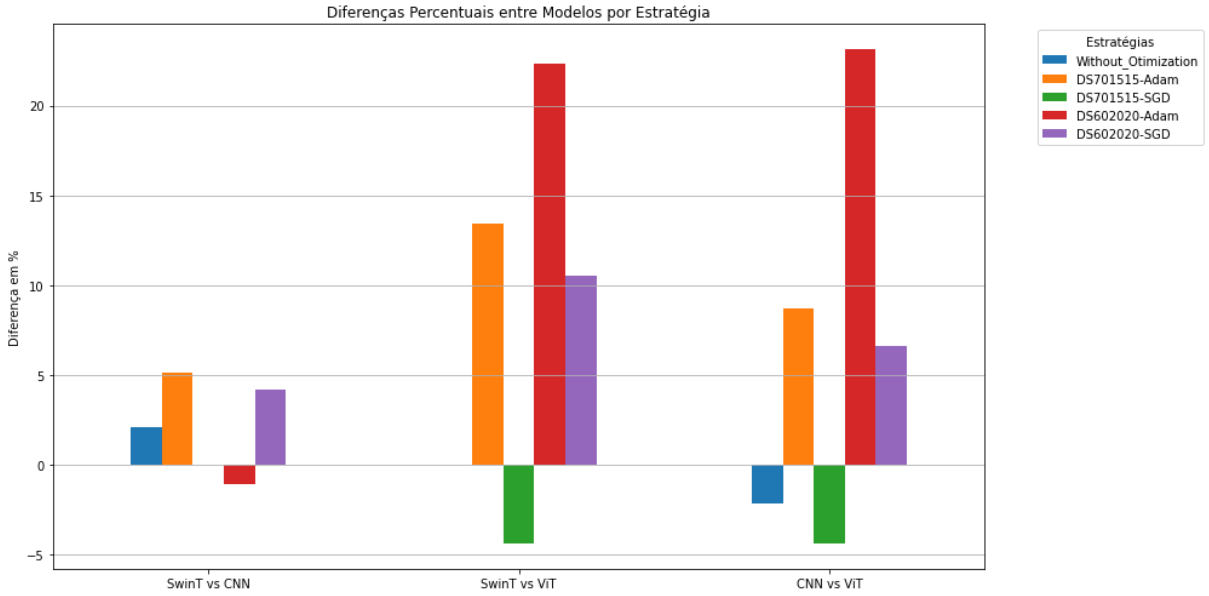


Fonte: o autor (2023).

Com base nos Gráficos 10 e 11 é possível verificar que as maiores vantagens obtidas por *Swin Transformer* e CNN sobre o ViT foram obtidas na estratégia de *split dataset* 60/20/20 utilizando otimizador Adam, seguido pela estratégia de *split dataset* 70/15/15 também com otimizador Adam. Embora o *Swin Transformer* tenha obtidos melhores resultados do que o CNN em três estratégias, a diferença não é tão grande, sendo a maior delas de 5%. Quanto aos dados de comparativos entre *Swin Transformer* e ViT a vantagem é mais relevante, pois o primeiro supera o segundo em mais de 20% na estratégia de *split dataset* 60/20/20 com otimizador Adam. Por outro lado, o ViT supero o *Swin Transformer* em quase 5% na estratégia de *split dataset* 70/15/15 com otimizador SDG.

Em seguida, no Gráfico 12, é possível verificar os resultados de AUC obtidos pelos modelos em cada *dataset*.

Gráfico 12 – Diferenças percentuais por estratégia para a AUC



Fonte: o autor (2023).

Uma análise detalhada mostra que:

1. SwinT *vs*. CNN:

- Without_Optimization: o modelo *SwinT* supera o modelo CNN com uma margem de 2.08 pontos percentuais na AUC.
- DS701515-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 5.15 pontos percentuais na AUC.
- DS701515-SGD: o modelo *SwinT* empata com o modelo CNN na AUC.
- DS602020-Adam: o modelo *SwinT* fica atrás do modelo CNN com uma diferença de 1.06 pontos percentuais na AUC.
- DS602020-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 4.21 pontos percentuais na AUC.

2. SwinT *vs*. ViT:

- Without_Optimization: o modelo *SwinT* supera o modelo ViT com uma margem de 0.00 pontos percentuais na AUC.
- DS701515-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 13.40 pontos percentuais na AUC.
- DS701515-SGD: o modelo *SwinT* fica atrás do modelo ViT com uma diferença de 4.40 pontos percentuais na AUC.

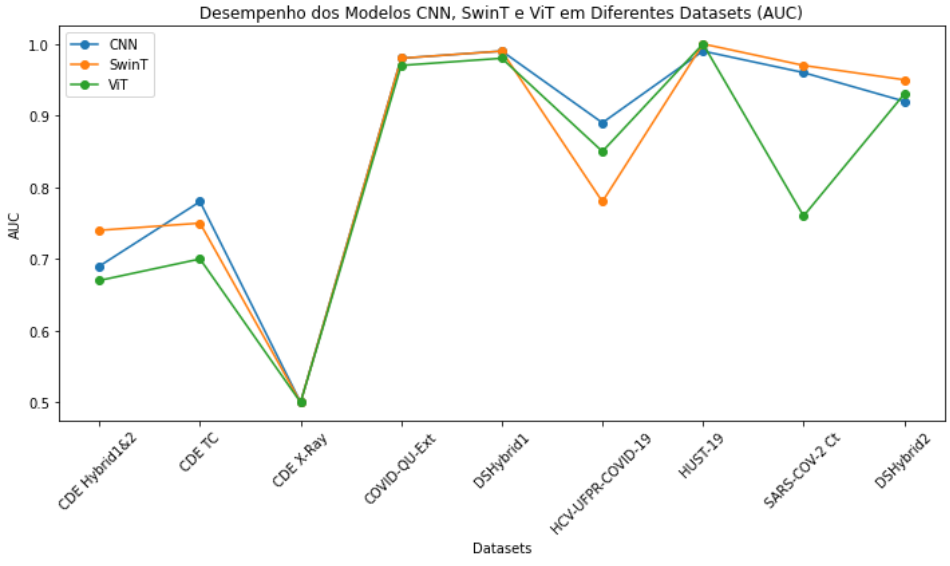
- DS602020-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 22.34 pontos percentuais na AUC.
- DS602020-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 10.53 pontos percentuais na AUC.

3. CNN *vs*. ViT:

- Without_Optimization: o modelo CNN supera o modelo ViT com uma margem de -2.13 pontos percentuais na AUC.
- DS701515-Adam: o modelo CNN supera o modelo ViT com uma margem de 8.70 pontos percentuais na AUC.
- DS701515-SGD: o modelo CNN fica atrás do modelo ViT com uma diferença de 4.40 pontos percentuais na AUC.
- DS602020-Adam: o modelo CNN supera o modelo ViT com uma margem de 23.16 pontos percentuais.
- DS602020-SGD: o modelo CNN supera o modelo ViT com uma margem de 6.59 pontos percentuais na AUC.

Essas comparações destacam as diferenças de desempenho entre os modelos em várias configurações, otimizadores e arquiteturas, com base nos novos dados. O Gráfico 13 demonstra que a AUC, no geral, se mantém elevada nos *datasets* originais e sofre considerável degradação nos *datasets* que empregaram estratégias de CDE. Quanto aos *datasets* com a estratégia híbrida é possível verificar que eles não sofreram tanta degradação AUC.

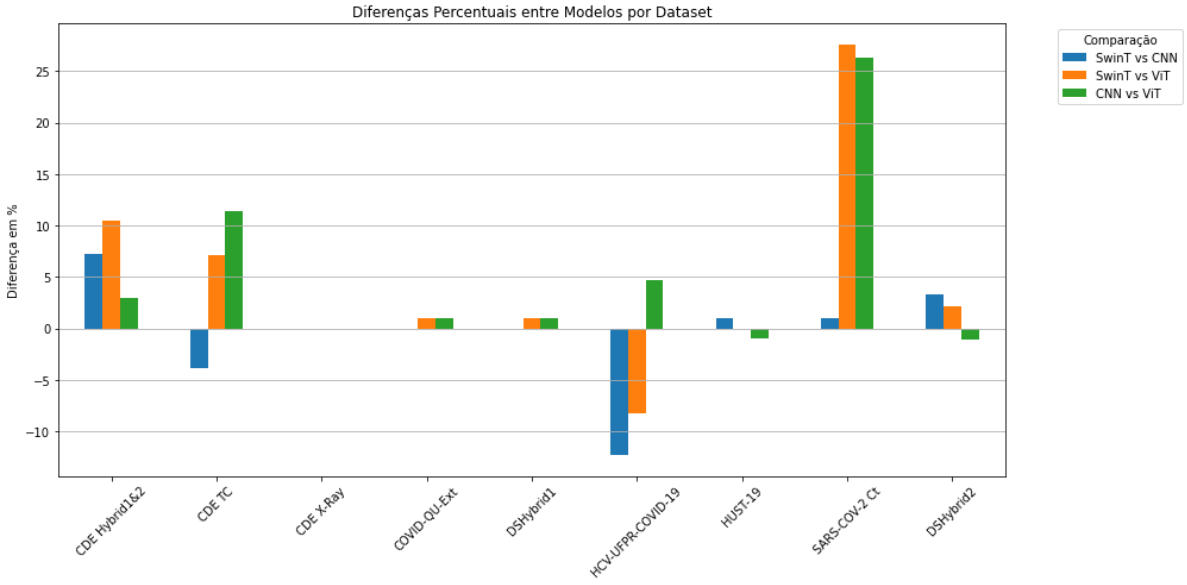
Gráfico 13 – Valores de AUC por modelo e *dataset*



Fonte: o autor (2023).

Para complementar a análise, o Gráfico 14 traz a diferença percentual dos valores obtidos para a AUC em cada modelo e *dataset*.

Gráfico 14 – Diferenças percentuais por *dataset* para a AUC



Fonte: o autor (2023).

Com base nos dados trazidos pelos Gráficos 13 e 14 exploramos e analisamos o desempenho dos três modelos nos nove conjuntos de dados com o objetivo de determinar como esses modelos se saíram em diferentes contextos e identificar suas vantagens e desvantagens. Os resultados estão compilados dessa forma:

a) CDE Hybrid1&2:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 7.25 pontos percentuais. Isso sugere que o *SwinT* é mais eficaz na classificação de dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 10.45 pontos percentuais. Isso indica que o *SwinT* é mais adequado para este conjunto de dados em comparação com o ViT.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 2.99 pontos percentuais. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

b) CDE TC:

1. *SwinT vs*. CNN: *SwinT* fica atrás do CNN com uma margem de 3.85 pontos percentuais. O CNN é mais eficiente na AUC para os dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 7.14 pontos percentuais. O SwinT demonstra um desempenho melhor em comparação com o ViT nesta configuração.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 11.43 pontos percentuais. Isso indica que o CNN é significativamente mais eficaz em relação ao ViT neste conjunto de dados.

c) CDE X-Ray, COVID-QU-Ext, DSHybrid1: para esses conjuntos de dados, todos os modelos apresentam um desempenho diferença percentual irrelevante. Nenhum dos modelos se destaca significativamente em termos de AUC de classificação nesses conjuntos de dados.

d) HCV-UFPR-COVID-19:

1. *SwinT vs*. CNN: *SwinT* fica atrás do CNN com uma margem de 12.36 pontos percentuais. O CNN supera o *SwinT* nesse contexto. Esse é o conjunto de dados onde o *SwinT* apresentou o pior desempenho em todas as métricas.

2. *SwinT vs*. ViT: *SwinT* fica atrás do ViT com uma margem de 8.24 pontos percentuais. O ViT supera o *SwinT* neste conjunto.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 4.71 pontos percentuais. O CNN é mais adequado para este conjunto de dados.

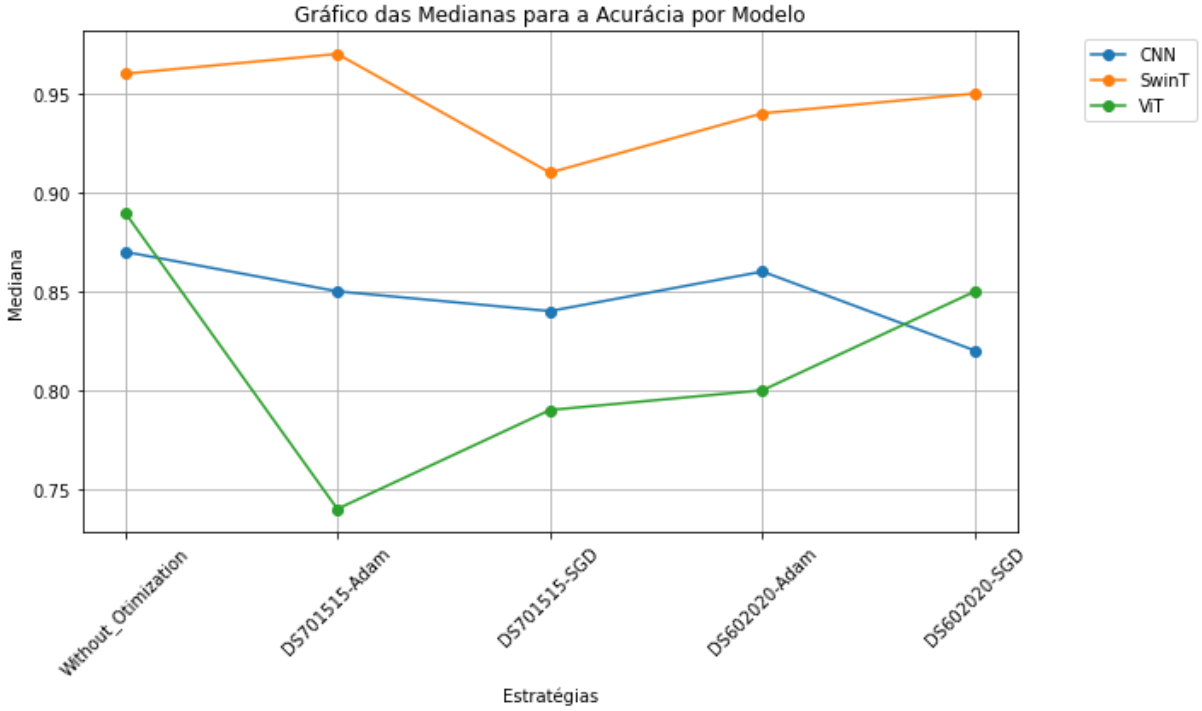
e) HUST-19, SARS-COV-2 Ct, DSHybrid2: o modelo *SwinT* supera o CNN e o ViT em todos esses conjuntos de dados. O *SwinT* demonstra uma vantagem consistente em termos de precisão de classificação. Cabe destacar que, no SARS-COV-2 Ct, o modelo ViT obteve desempenho bem abaixo do *SwinT* e CNN, 27.63 e 26.31 pontos percentuais, respectivamente.

Esses resultados da AUC evidenciam que o desempenho dos modelos varia significativamente de acordo com o conjunto de dados. O *SwinT* geralmente se destaca, mas existem casos em que o CNN ou o ViT podem ser mais adequados. Essas descobertas são valiosas para a seleção adequada de modelos em tarefas de classificação de diferentes tipos de dados médicos e de saúde pública.

4.4.2.2.2AED da Acurácia

O Gráfico 15 demonstra que a acurácia teve diferentes resultados nas diferentes estratégias e modelos. Pode-se verificar que o modelo *Swin Transformer* obteve valores superiores aos do CNN e ViT. Embora o ViT tenha superado a CNN em duas estratégias, ele ficou em último lugar devido a ter obtidos valores inferiores aos do CNN nas outras três estratégias.

Gráfico 15 – Medianas da acurácia para as 15 diferentes estratégias



Fonte: o autor (2023).

Com base nos Gráficos 14 e 15 é possível verificar que a vantagem obtida pelo *Swin Transformer* sobre o CNN e o ViT é expressiva e consistente, tendo sido obtida em todas as estratégias. Uma análise detalhada nos mostra que:

1. *SwinT vs*. CNN:

- Without_Optimization: o modelo *SwinT* supera o modelo CNN em desempenho com uma margem considerável de 9.38 pontos percentuais. Isso sugere que o *SwinT* é mais eficaz na tarefa em questão em comparação com o CNN.
- DS701515-Adam: o modelo *SwinT* também supera o modelo CNN com a otimização do Adam, porém com uma melhoria ainda maior de

12.37 pontos percentuais. Isso indica que o uso de Adam como otimizador beneficia ambos os modelos, mas o *SwinT* mantém sua vantagem.

- DS701515-SGD: com o otimizador SGD, o modelo *SwinT* continua a superar o modelo CNN, mas a diferença é menor, com uma margem de 7.69 pontos percentuais. O SGD parece ser menos eficaz do que o Adam para ambos os modelos, mas o *SwinT* ainda se destaca.
- DS602020-Adam: neste caso, o modelo *SwinT* mantém uma vantagem sobre o CNN, mas com uma diferença menor de 8.51 pontos percentuais. O uso de uma estratégia de *split dataset* diferente (60/20/20) não afeta drasticamente as diferenças entre os modelos.
- DS602020-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 13.68 pontos percentuais.

2. SwinT *vs*. ViT:

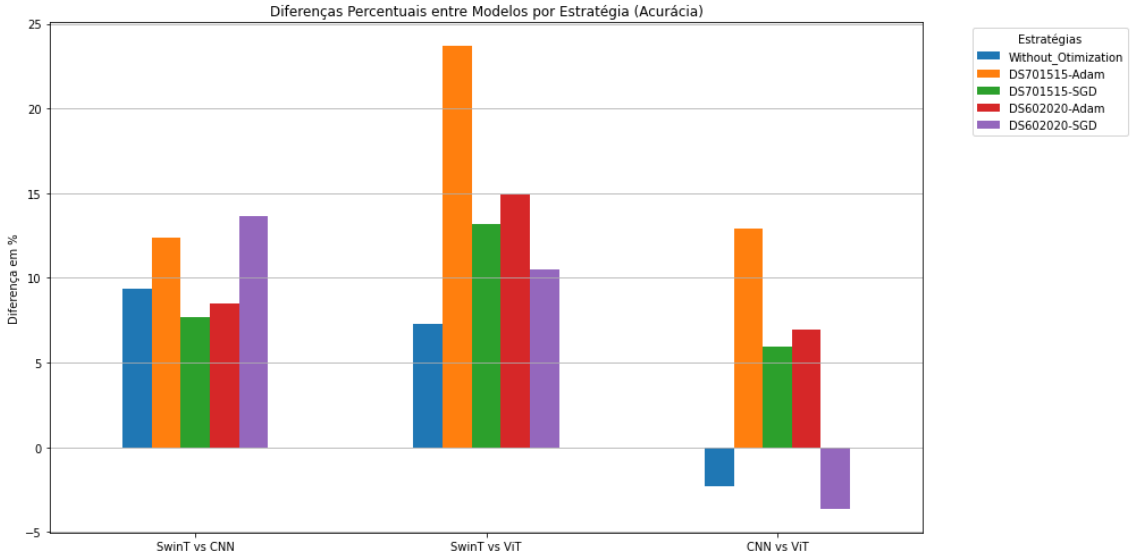
- Without_Optimization: o modelo *SwinT* supera o modelo ViT com uma margem de 7.29 pontos percentuais. Isso sugere que, no contexto específico deste problema, o *SwinT* tem uma vantagem inicial em relação ao ViT.
- DS701515-Adam: o modelo ViT com otimização do Adam melhora seu desempenho e se aproxima do *SwinT*, mas ainda fica atrás, com uma diferença de 23.71 versus 12.37 pontos percentuais. Adam parece ser benéfico para ambos os modelos, mas o *SwinT* mantém uma liderança considerável.
- DS701515-SGD: o modelo ViT com SGD ainda fica atrás do *SwinT*, mas com uma diferença menor de 13.19 versus 7.69 pontos percentuais. O uso do SGD tem um impacto mais significativo no desempenho do ViT, mas o *SwinT* ainda é superior.
- DS602020-Adam: com o uso do Adam e a mudança na estratégia de *split dataset*, o modelo ViT reduz ainda mais a diferença para 14.89 versus 8.51 pontos percentuais, mas o *SwinT* continua liderando.
- DS602020-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 10.53 pontos percentuais.

3. CNN *vs*. ViT:

- Without_Optimization: inicialmente, o modelo CNN é superado pelo modelo ViT com uma margem de 2.30 pontos percentuais. No entanto, a diferença é relativamente pequena e pode não ser estatisticamente significativa.
- DS701515-Adam: com o otimizador Adam, a diferença entre o modelo CNN e o modelo ViT aumenta para 12.94 pontos percentuais, com vantagem para o CNN. Isso sugere que o Adam beneficia mais o CNN nesta configuração.
- DS701515-SGD: com o SGD, o modelo CNN mantém sua vantagem sobre o ViT, mas a diferença diminui para 5.95 pontos percentuais. O SGD parece beneficiar mais o CNN em comparação com o ViT.
- DS602020-Adam: com a mudança na estratégia de *split dataset* para 60/20/20 e o otimizador Adam, a diferença aumenta para 6.98 pontos percentuais, ainda favorecendo o CNN.
- DS602020-SGD: o modelo CNN é superado pelo modelo ViT com uma margem de 3.66 pontos percentuais na acurácia.

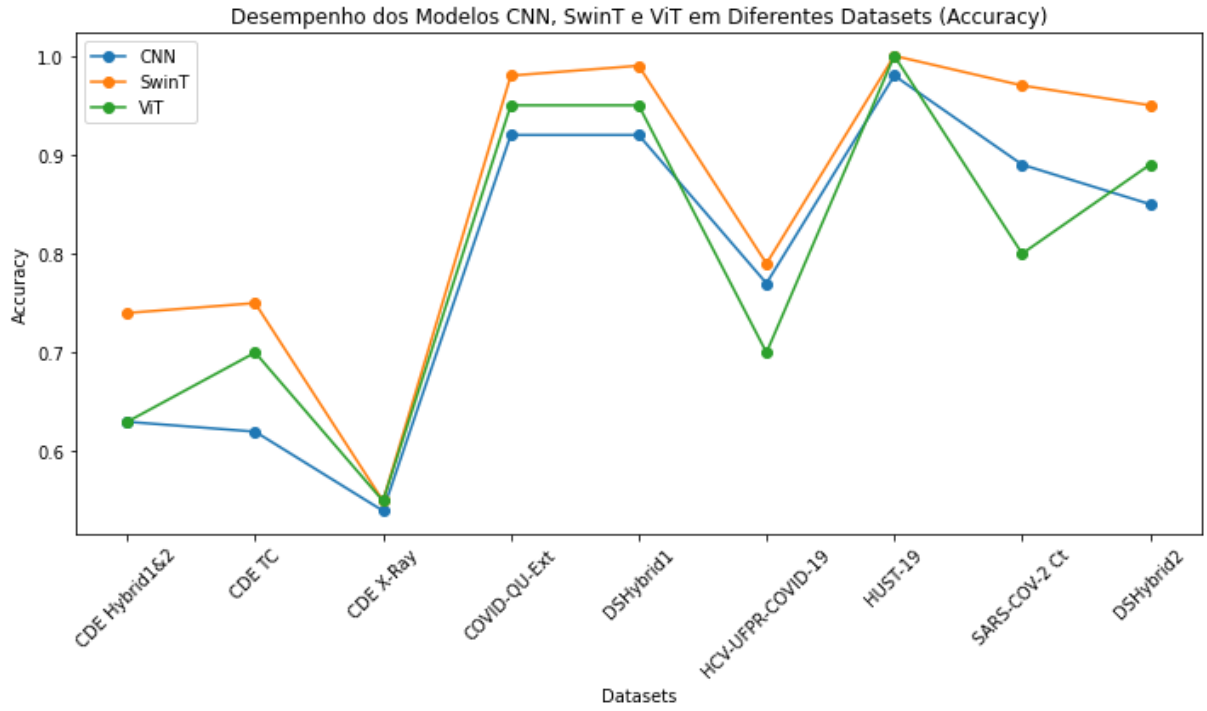
Em resumo, as comparações mostram que o modelo *SwinT* tende a superar tanto o CNN quanto o ViT em diferentes configurações e otimizadores, com algumas variações nas diferenças de desempenho. O CNN também supera o ViT em algumas configurações, especialmente quando se usa o otimizador Adam.

Gráfico 16 – Diferenças percentuais por estratégia para a acurácia



Fonte: o autor (2023).

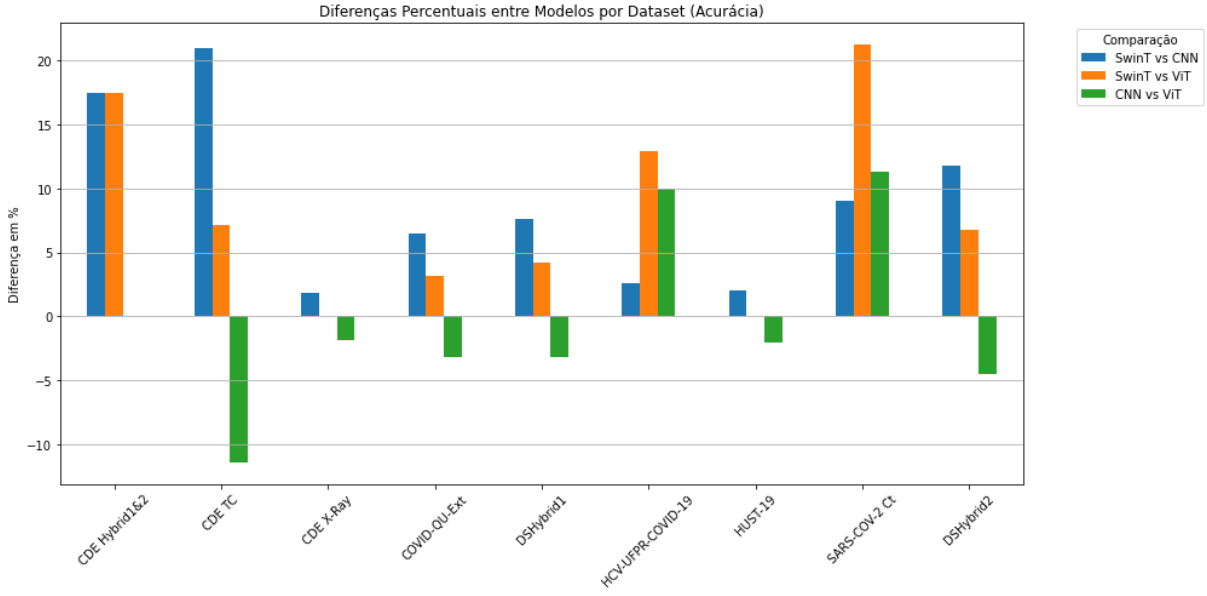
Gráfico 17 – Medianas da Acurácia para os nove diferentes *datasets*



Fonte: o autor (2023).

Para complementar a análise, o Gráfico 18 traz a diferença percentual dos valores obtidos para a acurácia em cada modelo e *dataset*.

Gráfico 18 – Diferenças percentuais por *dataset* para a Acurácia



Fonte: o autor (2023).

Com base nos dados trazidos pelos Gráficos 16, 17 e 18, exploramos e analisamos o desempenho dos três modelos nos nove conjuntos de dados com

o objetivo de determinar como esses modelos se saíram em diferentes contextos e identificar suas vantagens e desvantagens. Os resultados estão compilados dessa forma:

a) CDE Hybrid1&2:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 17.46 pontos percentuais em termos de acurácia. Isso indica que o *SwinT* é mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 17.46 pontos percentuais em termos de acurácia. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a acurácia.

3. CNN *vs*. ViT: o CNN empata com o ViT com uma margem em termos de acurácia.

b) CDE TC:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 20.97 pontos percentuais em termos de acurácia. Isso indica que o SwinT é mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 7.14 pontos percentuais em termos de acurácia. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a acurácia.

3. CNN *vs*. ViT: o CNN é superado pelo ViT com uma margem de 11.42 pontos percentuais em termos de acurácia. Isso mostra que o ViT é mais competitivo em relação ao CNN neste contexto.

c) CDE X-Ray, COVID-QU-Ext, DSHybrid1: Para esses conjuntos de dados, todos os modelos apresentam desempenho com diferença em torno de 5% em termos de acurácia. Ainda assim, há liderança do SwinT nesses três conjuntos de dados.

d) HCV-UFPR-COVID-19:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 2.60 pontos percentuais em termos de acurácia.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 12.86 pontos percentuais em termos de acurácia.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de dez pontos percentuais em termos de acurácia. O CNN é mais adequado para este conjunto de dados.

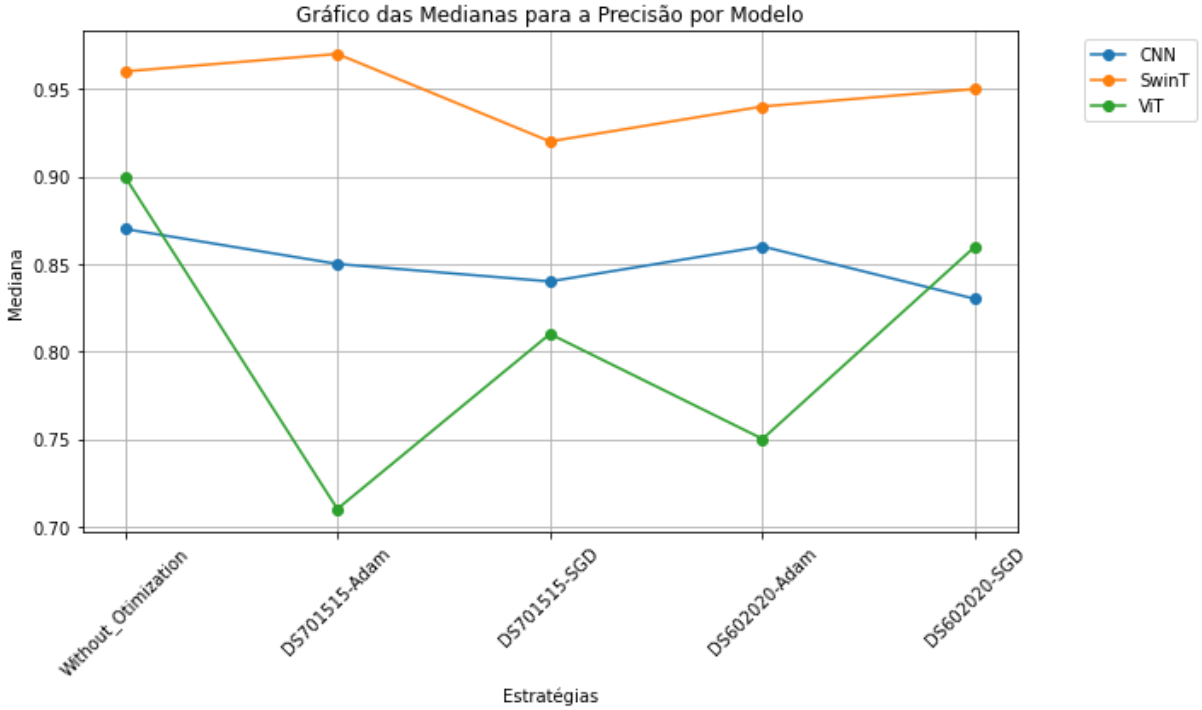
e) HUST-19, SARS-COV-2 Ct, DSHybrid2: nesses conjuntos de dados, o *SwinT* supera o CNN e o ViT em termos de acurácia, com margens variadas. Pode-se destacar que a liderança do *SwinT* para a acurácia é consistente nesses conjuntos de dados.

Em resumo, esses resultados destacam a importância de selecionar cuidadosamente o modelo de rede neural com base no conjunto de dados e na métrica de avaliação desejada. O *SwinT* frequentemente demonstra um desempenho superior em acurácia, mas podem existir casos em que o CNN ou o ViT podem ser mais adequados, dependendo do contexto da tarefa de classificação.

4.4.2.2.3 AED da Precisão

O Gráfico 19 demonstra que a precisão teve diferentes resultados nas diferentes estratégias e modelos. Pode-se verificar que o modelo *Swin Transformer* obteve valores superiores aos do CNN e ViT. Embora o ViT tenha superado a CNN em duas estratégias, ele ficou em último lugar devido a ter obtidos valores inferiores aos do CNN nas outras três estratégias.

Gráfico 19 – Medianas da precisão para as 15 diferentes estratégias



Fonte: o autor (2023).

Com base nos Gráficos 18 e 19, é possível verificar que a vantagem obtida pelo *Swin Transformer* sobre o CNN e o ViT é expressiva e consistente, tendo sido obtida em todas as estratégias. Uma análise detalhada nos mostra que:

1. *SwinT vs*. CNN:

- Without_Optimization: o modelo *SwinT* supera o modelo CNN com uma margem de 9.38 pontos percentuais na precisão.
- DS701515-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 12.37 pontos percentuais na precisão.
- DS701515-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 8.70 pontos percentuais na precisão.
- DS602020-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 8.51 pontos percentuais na precisão.
- DS602020-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 12.63 pontos percentuais na precisão.

2. SwinT *vs*. ViT:

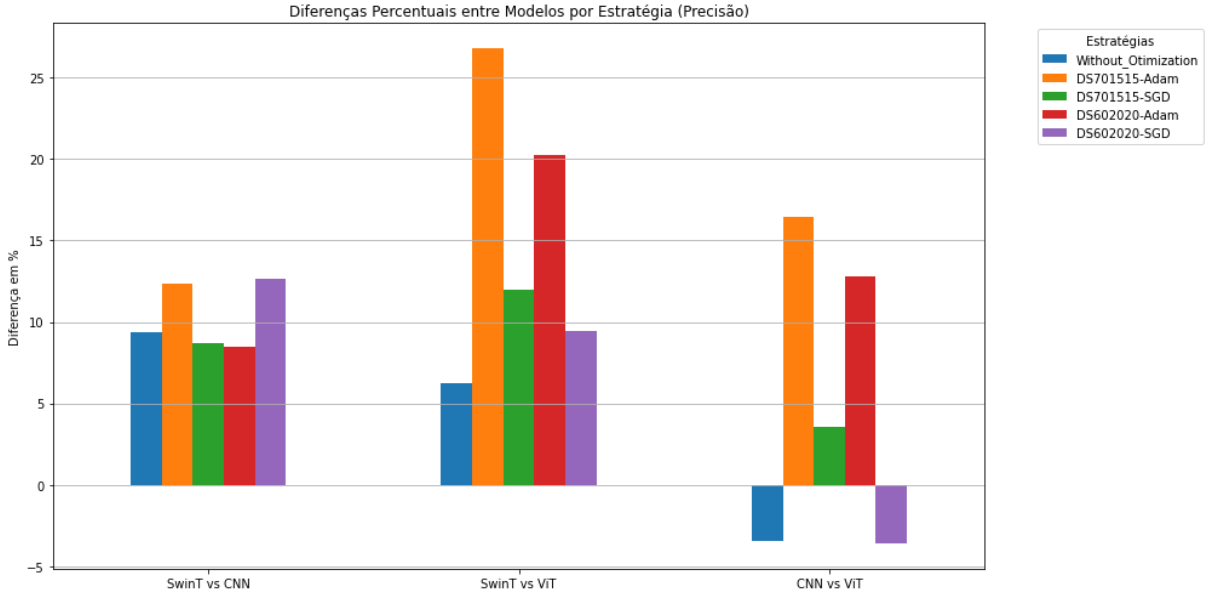
- Without_Optimization: o modelo *SwinT* supera o modelo ViT com uma margem de 6.25 pontos percentuais na precisão.
- DS701515-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 26.80 pontos percentuais na precisão.
- DS701515-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 11.96 pontos percentuais na precisão.
- DS602020-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 20.21 pontos percentuais na precisão.
- DS602020-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 9.47 pontos percentuais na precisão.

3. CNN *vs*. ViT:

- Without_Optimization: o modelo CNN fica atrás o modelo ViT com uma margem de 3.45 pontos percentuais na precisão.
- DS701515-Adam: o modelo CNN supera o modelo ViT com uma margem de 16.47 pontos percentuais na precisão.
- DS701515-SGD: o modelo CNN supera o modelo ViT com uma margem de 3.57 pontos percentuais na precisão.

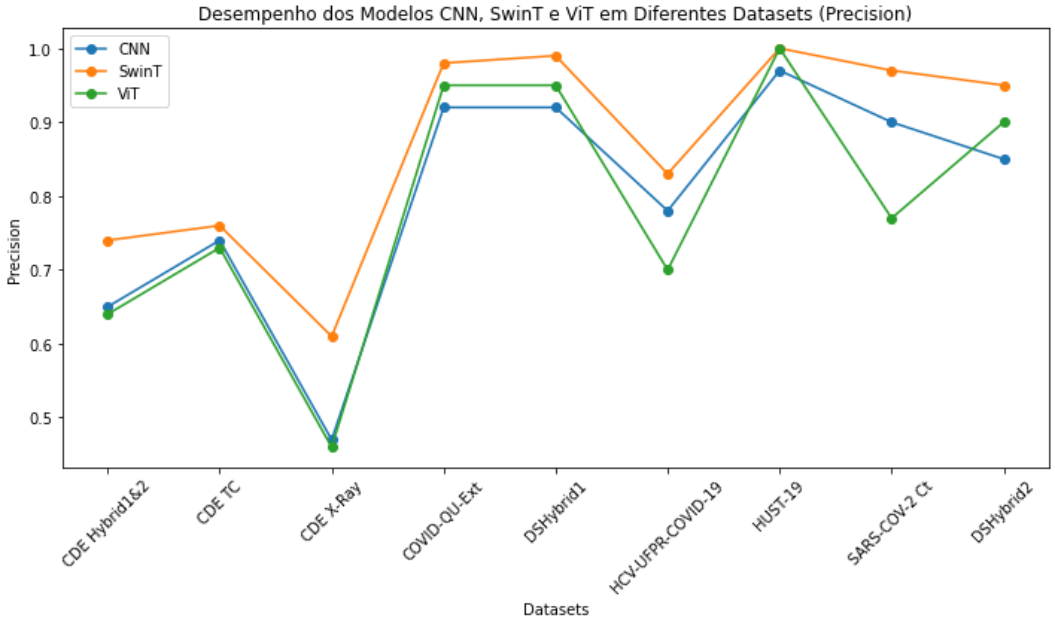
- DS602020-Adam: o modelo CNN supera o modelo ViT com uma margem de 12.79 pontos percentuais na precisão.
- DS602020-SGD: o modelo CNN supera o modelo ViT com uma margem de -3.61 pontos percentuais na precisão.

Gráfico 20 – Diferenças percentuais por estratégia para a Precisão



Fonte: o autor (2023).

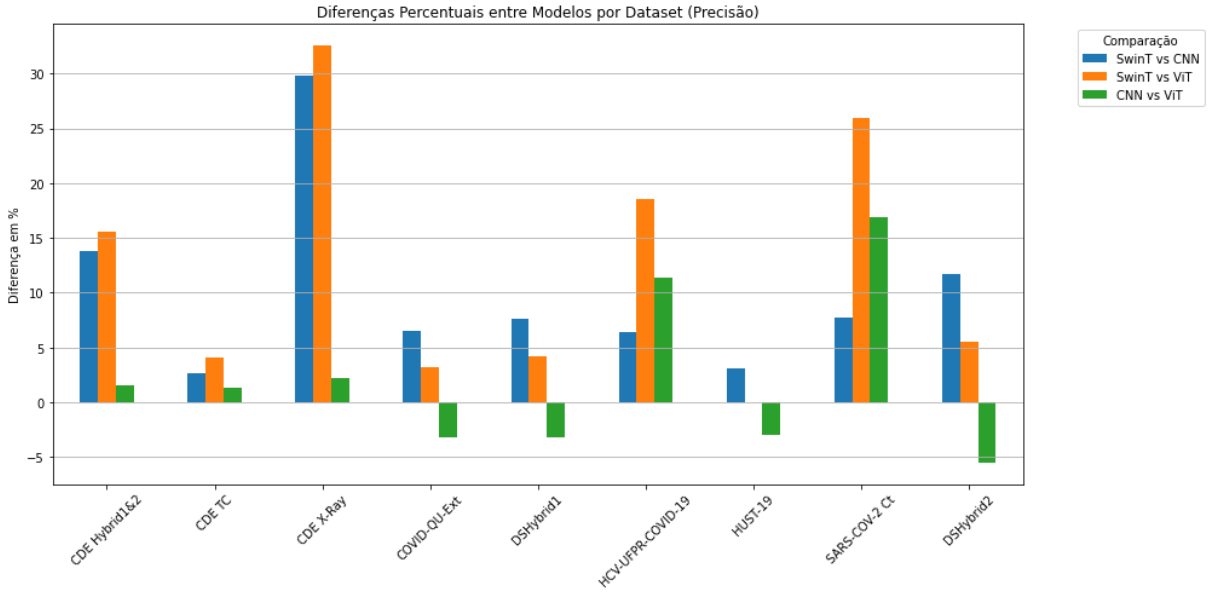
Gráfico 21 – Valores de precisão por modelo e *dataset*



Fonte: o autor (2023).

Para complementar a análise, o Gráfico 22 traz a diferença percentual dos valores obtidos para a precisão em cada modelo e *dataset*.

Gráfico 22 – Diferenças percentuais por *dataset* para a Precisão



Fonte: o autor (2023).

Com base nos dados trazidos nos Gráficos 21 e 22, exploramos e analisamos o desempenho dos três modelos nos nove conjuntos de dados com o objetivo de determinar como esses modelos se saíram em diferentes contextos e identificar suas vantagens e desvantagens. Os resultados estão compilados dessa forma:

a) CDE Hybrid1&2:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 13.85 pontos percentuais. Isso indica que o *SwinT* é mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 15.62 pontos percentuais. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a precisão.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 1.56 pontos percentuais. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

b) CDE TC:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 2.70 pontos percentuais. Isso indica que o *SwinT* é mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 4.10 pontos percentuais. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a precisão.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 1.37 pontos percentuais. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

c) CDE *X-Ray*:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 29.78 pontos percentuais. Isso indica que o SwinT é significativamente mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 32.61 pontos percentuais. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a precisão.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 2.17 pontos percentuais. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

d) COVID-QU-Ext, DSHybrid1: para esses conjuntos de dados, o *SwinT* supera os outros modelos em termos de precisão, com margens variadas.

e) HCV-UFPR-COVID-19:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 6.41 pontos percentuais. Isso indica que o *SwinT* é mais preciso na classificação dos dados deste conjunto.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 18.57 pontos percentuais. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a precisão.

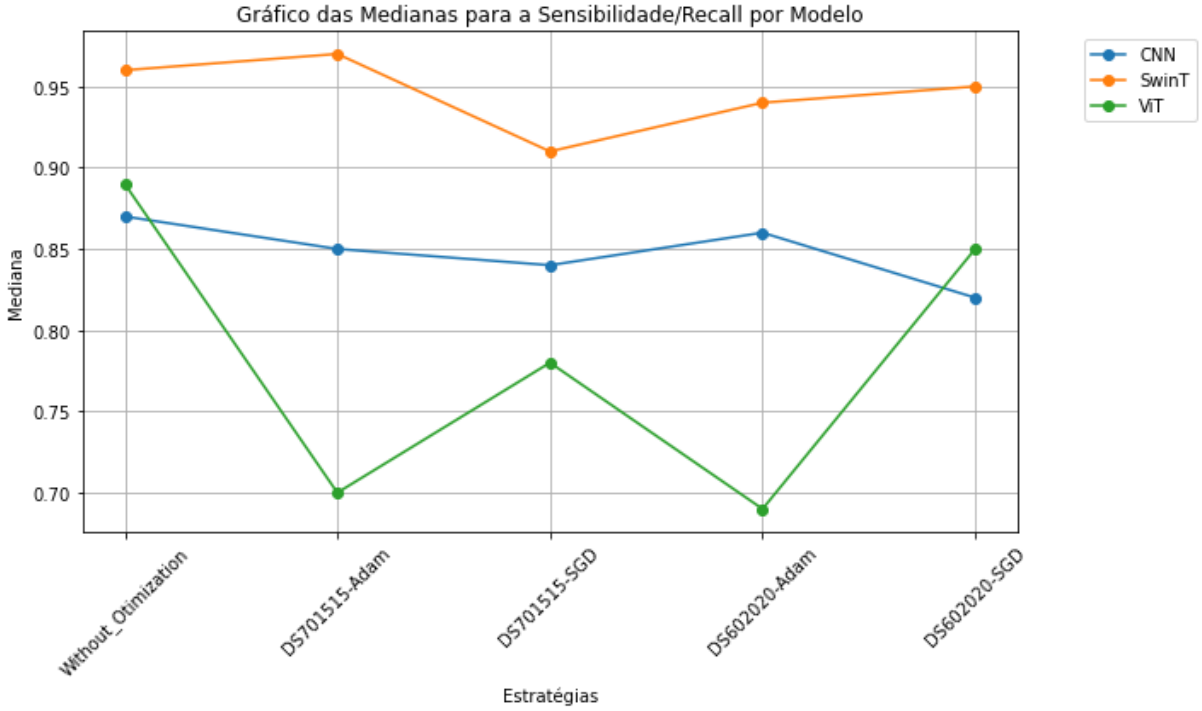
3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 11.43 pontos percentuais. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

f) HUST-19, SARS-COV-2 Ct, DSHybrid2: nesses conjuntos de dados, o *SwinT* supera consistentemente os outros modelos em termos de precisão, com margens variadas.

4.4.2.2.4 AED da Sensibilidade/Recall

O Gráfico 23 demonstra que a sensibilidade teve diferentes resultados nas diferentes estratégias e modelos. Pode-se verificar que o modelo *Swin Transformer* obteve valores superiores aos do CNN e ViT. Embora o ViT tenha superado a CNN em duas estratégias, ele ficou em último lugar devido a ter obtidos valores inferiores aos do CNN nas outras três estratégias.

Gráfico 23 – Medianas da sensibilidade para as 15 diferentes estratégias



Fonte: o autor (2023).

Com base nos Gráficos 22 e 23, é possível verificar que a vantagem obtida pelo *Swin Transformer* sobre o CNN e o ViT é expressiva e consistente, tendo sido obtida em todas as estratégias. Uma análise detalhada nos mostra que:

1. *SwinT vs*. CNN:

- Without_Optimization: o modelo *SwinT* supera o modelo CNN com uma margem de 9.38 pontos percentuais na sensibilidade.
- DS701515-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 12.37 pontos percentuais na sensibilidade.
- DS701515-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 7.69 pontos percentuais na sensibilidade.

- DS602020-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 8.51 pontos percentuais na sensibilidade.
- DS602020-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 13.68 pontos percentuais na sensibilidade.

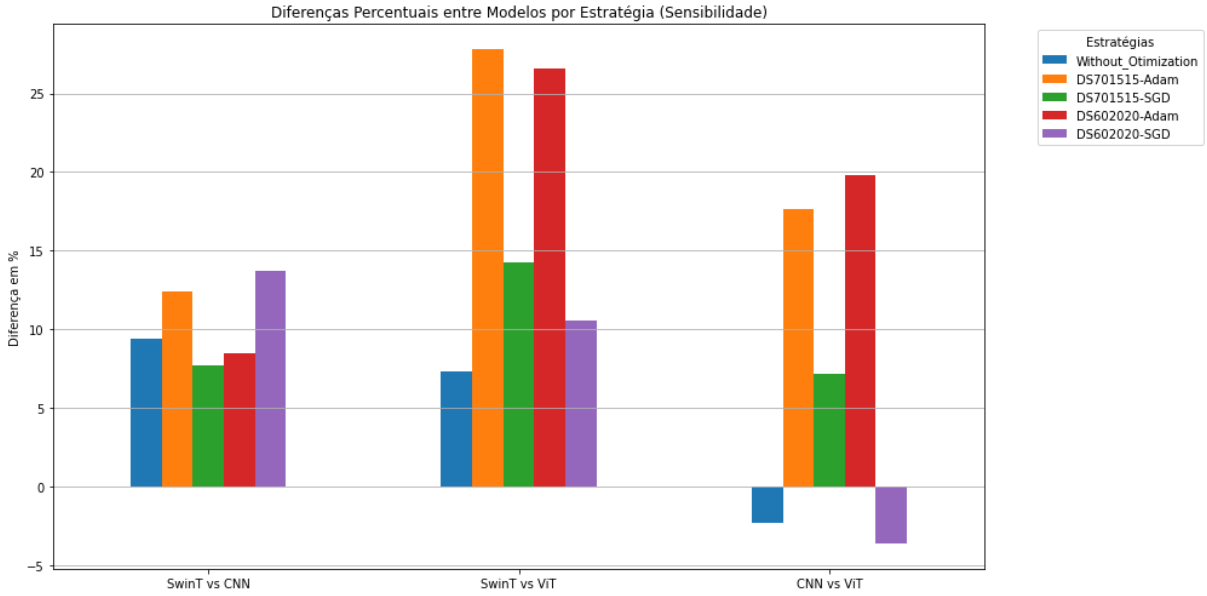
2. *SwinT vs*. ViT:

- Without_Optimization: o modelo *SwinT* supera o modelo ViT com uma margem de 7.29 pontos percentuais na sensibilidade.
- DS701515-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 27.84 pontos percentuais na sensibilidade.
- DS701515-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 14.29 pontos percentuais na sensibilidade.
- DS602020-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 26.60 pontos percentuais na sensibilidade.
- DS602020-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 10.53 pontos percentuais na sensibilidade.

3. CNN *vs*. ViT:

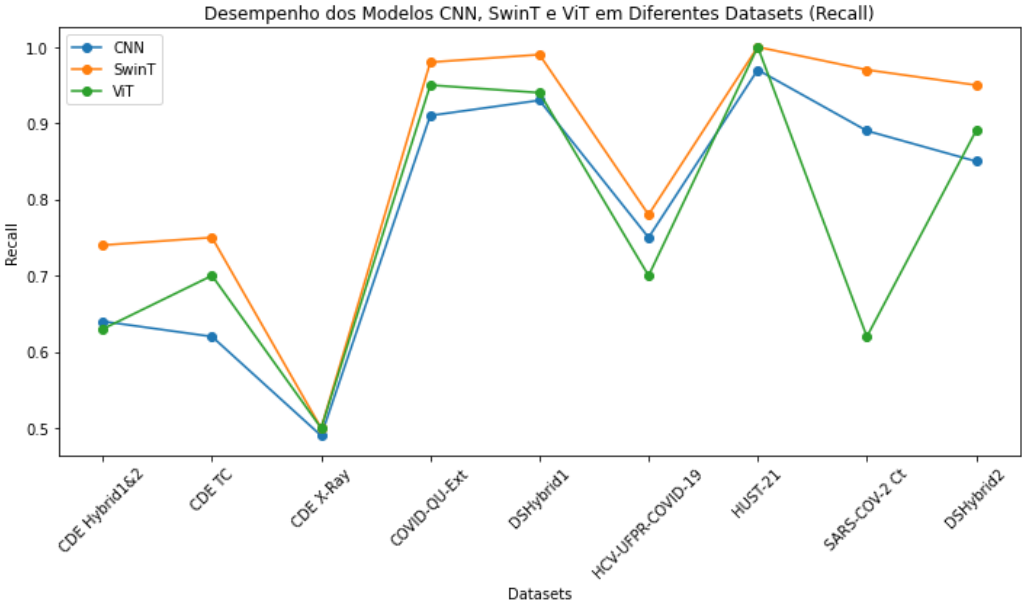
- Without_Optimization: o modelo CNN fica atrás do modelo ViT por uma margem de 2.30 pontos percentuais na sensibilidade.
- DS701515-Adam: o modelo CNN supera o modelo ViT com uma margem de 17.65 pontos percentuais na sensibilidade.
- DS701515-SGD: o modelo CNN supera o modelo ViT com uma margem de 7.14 pontos percentuais na sensibilidade.
- DS602020-Adam: o modelo CNN supera o modelo ViT com uma margem de 19.77 pontos percentuais na sensibilidade.
- DS602020-SGD: o modelo CNN fica atrás do modelo ViT por uma margem de 3.66 pontos percentuais na sensibilidade.

Gráfico 24 – Diferenças percentuais por estratégia para a sensibilidade



Fonte: o autor (2023).

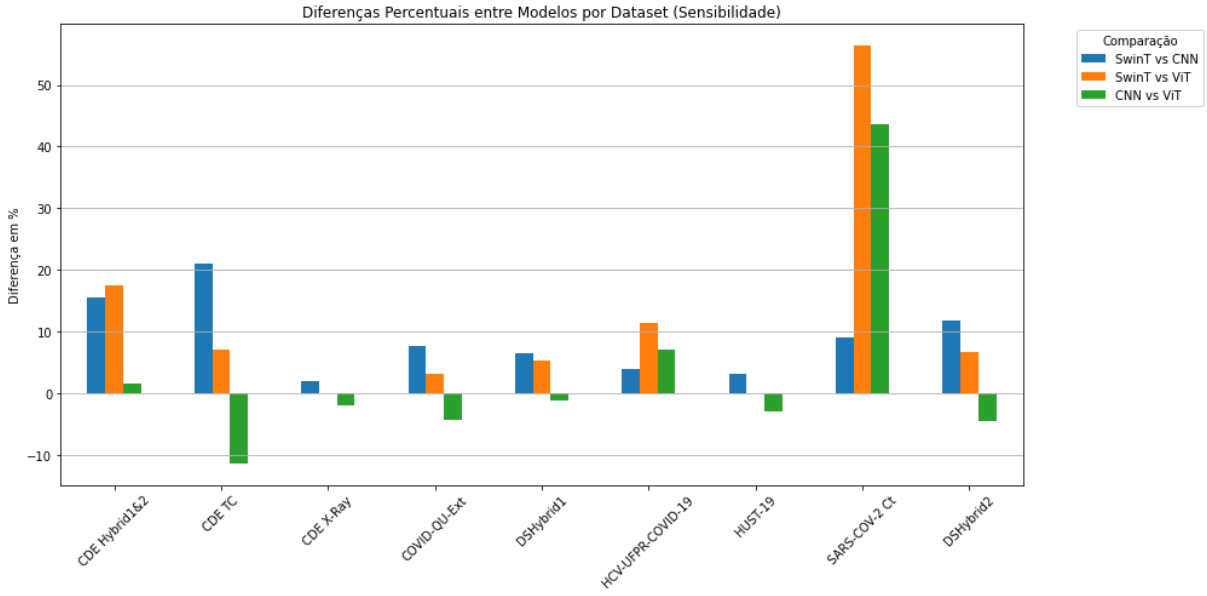
Gráfico 25 – Valores de sensibilidade por modelo e *dataset*



Fonte: o autor (2023).

Para complementar a análise, o Gráfico 26 traz a diferença percentual dos valores obtidos para a sensibilidade em cada modelo e *dataset*.

Gráfico 26 – Diferenças percentuais por *dataset* para a sensibilidade



Fonte: o autor (2023).

Com base nos dados trazidos pelos Gráficos 24, 25 e 26, exploramos e analisamos o desempenho dos três modelos nos nove conjuntos de dados com o objetivo de determinar como esses modelos se saíram em diferentes contextos e identificar suas vantagens e desvantagens. Os resultados estão compilados dessa forma:

a) CDE Hybrid1&2:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 15.63 pontos percentuais em termos de sensibilidade. Isso indica que o *SwinT* é melhor na identificação de positivos verdadeiros nesse conjunto de dados.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 17.46 pontos percentuais em termos de sensibilidade. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a sensibilidade.

3. CNN *vs*. ViT: o CNN supera o ViT com uma margem de 1.59 pontos percentuais em termos de sensibilidade. Isso mostra que o CNN é mais competitivo em relação ao ViT neste contexto.

b) CDE TC:

1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 20.97 pontos percentuais em termos de sensibilidade. Isso indica que o

SwinT é melhor na identificação de positivos verdadeiros neste conjunto de dados.

2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 7.14 pontos percentuais em termos de sensibilidade. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar a sensibilidade.

3. CNN *vs*. ViT: o CNN é superado pelo ViT com uma margem de 11.43 pontos percentuais em termos de sensibilidade. Isso mostra que o CNN é significativamente menos competitivo em relação ao ViT neste contexto.

c) CDE *X-Ray*, COVID-QU-Ext, DSHybrid1, HCV-UFPR-COVID-19, HUST-19: para esses conjuntos de dados, os resultados variam, com algumas métricas indicando vantagem para o *SwinT* ou ViT, dependendo do contexto.

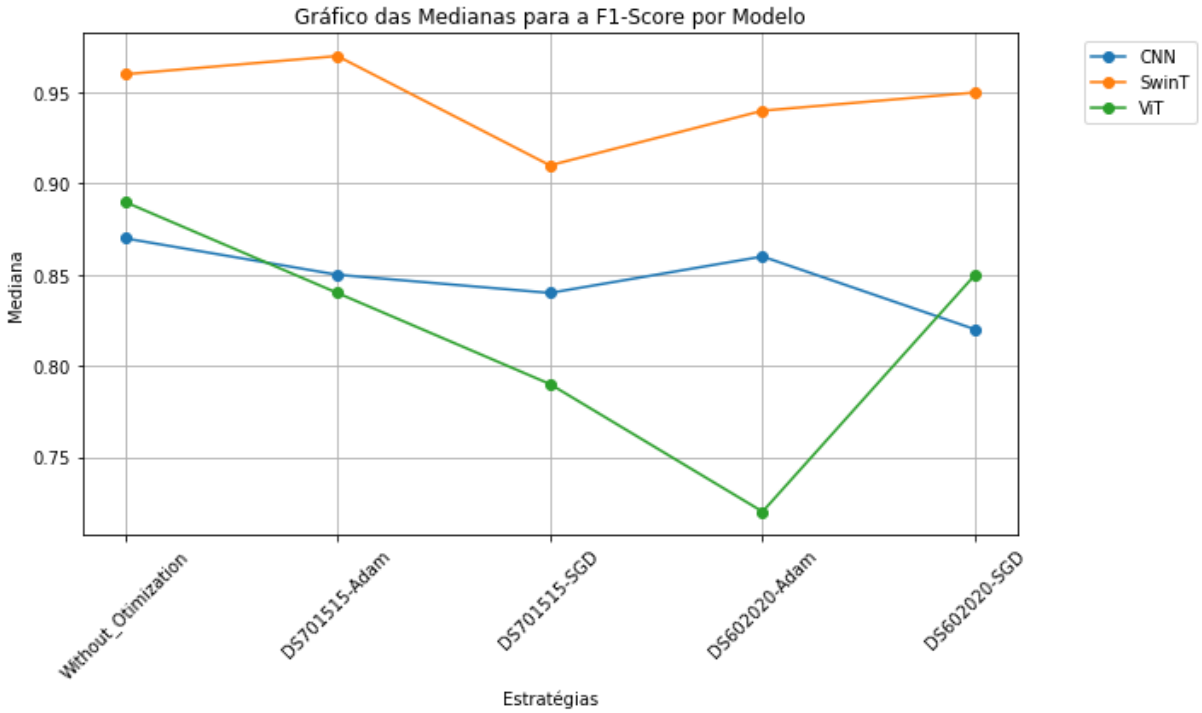
d) SARS-COV-2 Ct, DSHybrid2: nesses conjuntos de dados, o *SwinT* supera o CNN e o ViT em termos de sensibilidade, com margens variadas. Cabe destacar aqui a maior margem de diferença registrada entre os modelos onde o *SwinT* superou o ViT por 56.45 pontos percentuais.

Em resumo, esses resultados destacam a importância de selecionar cuidadosamente o modelo de rede neural com base no conjunto de dados e na métrica de avaliação desejada. O *SwinT* frequentemente demonstra um desempenho superior em sensibilidade.

4.4.2.2.5 AED da F1-Score

O Gráfico 27 demonstra que a F1-Score teve diferentes resultados nas diferentes estratégias e modelos. Pode-se verificar que o modelo *Swin Transformer* obteve valores superiores aos do CNN e ViT. Embora o ViT tenha superado a CNN em duas estratégias, ele ficou em último lugar devido a ter obtidos valores inferiores aos do CNN nas outras três estratégias.

Gráfico 27 – Medianas da F1-Score para as 15 diferentes estratégias



Fonte: o autor (2023).

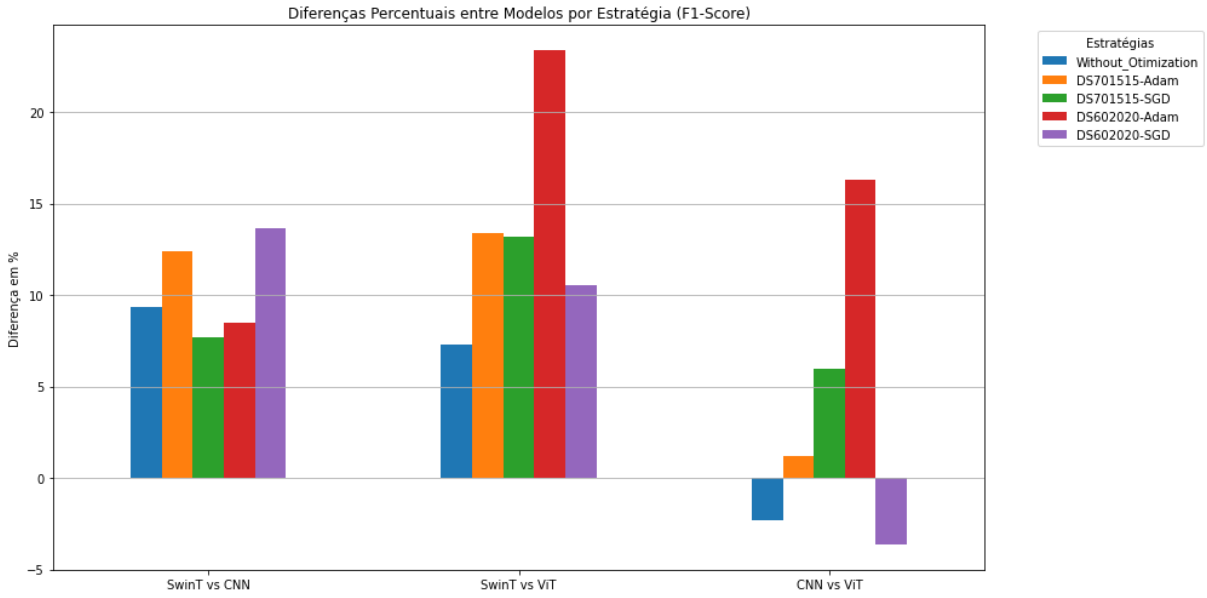
Com base nos Gráficos 26 e 27, é possível verificar que a vantagem obtida pelo *Swin Transformer* sobre o CNN e o ViT é expressiva e consistente, tendo sido obtida em todas as estratégias. Uma análise detalhada nos mostra que:

1. *SwinT vs*. CNN:
    - Without_Optimization: o modelo *SwinT* supera o modelo CNN com uma margem de 9.38 pontos percentuais na F1-Score.
    - DS701515-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 12.37 pontos percentuais na F1-Score.
    - DS701515-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 7.69 pontos percentuais na F1-Score.
    - DS602020-Adam: o modelo *SwinT* supera o modelo CNN com uma margem de 8.51 pontos percentuais na F1-Score.
    - DS602020-SGD: o modelo *SwinT* supera o modelo CNN com uma margem de 13.68 pontos percentuais na F1-Score.
2. *SwinT vs*. ViT:
    - Without_Optimization: o modelo *SwinT* supera o modelo ViT com uma margem de 7.29 pontos percentuais na F1-Score.

- DS701515-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 13.40 pontos percentuais na F1-Score.
- DS701515-SGD: o modelo *SwinT* supera o modelo ViT com uma margem de 13.19 pontos percentuais na F1-Score.
- DS602020-Adam: o modelo *SwinT* supera o modelo ViT com uma margem de 23.40 pontos percentuais na F1-Score.
- DS602020-SGD: O modelo *SwinT* supera o modelo ViT com uma margem de 10.53 pontos percentuais na F1-Score.

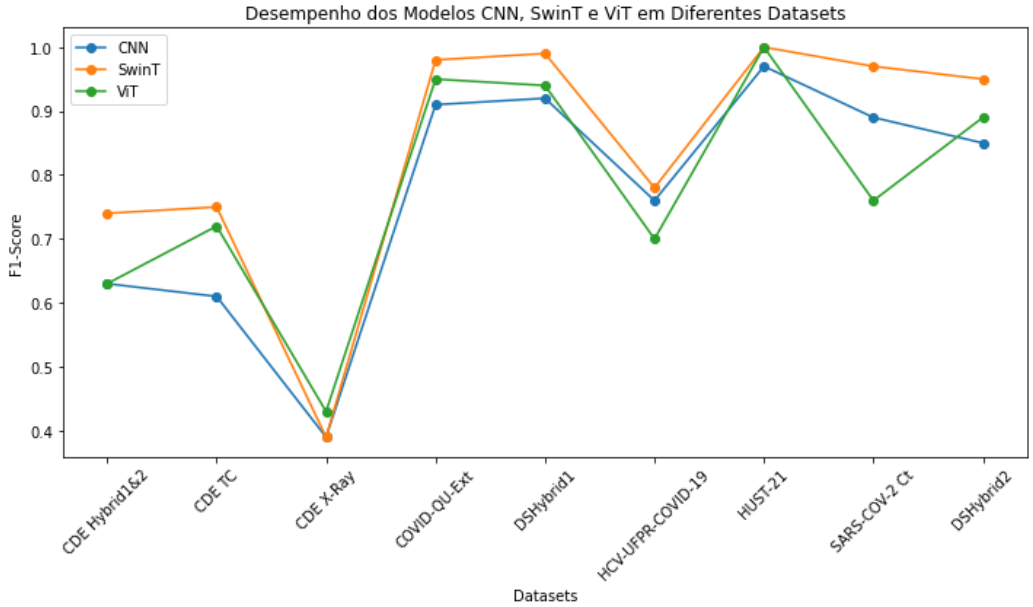
3. CNN *vs*. ViT:
   - Without_Optimization: o modelo CNN fica atrás do modelo ViT por uma margem de 2.30 pontos percentuais na F1-Score.
   - DS701515-Adam: o modelo CNN supera o modelo ViT com uma margem de 1.18 pontos percentuais na F1-Score.
   - DS701515-SGD: o modelo CNN supera o modelo ViT com uma margem de 5.95 pontos percentuais na F1-Score.
   - DS602020-Adam: o modelo CNN supera o modelo ViT com uma margem de 16.28 pontos percentuais na F1-Score.
   - DS602020-SGD: o modelo CNN fica atrás do modelo ViT por uma margem de 3.66 pontos percentuais na F1-Score.

Gráfico 28 – Diferenças percentuais por estratégia para a F1-Score



Fonte: o autor (2023).

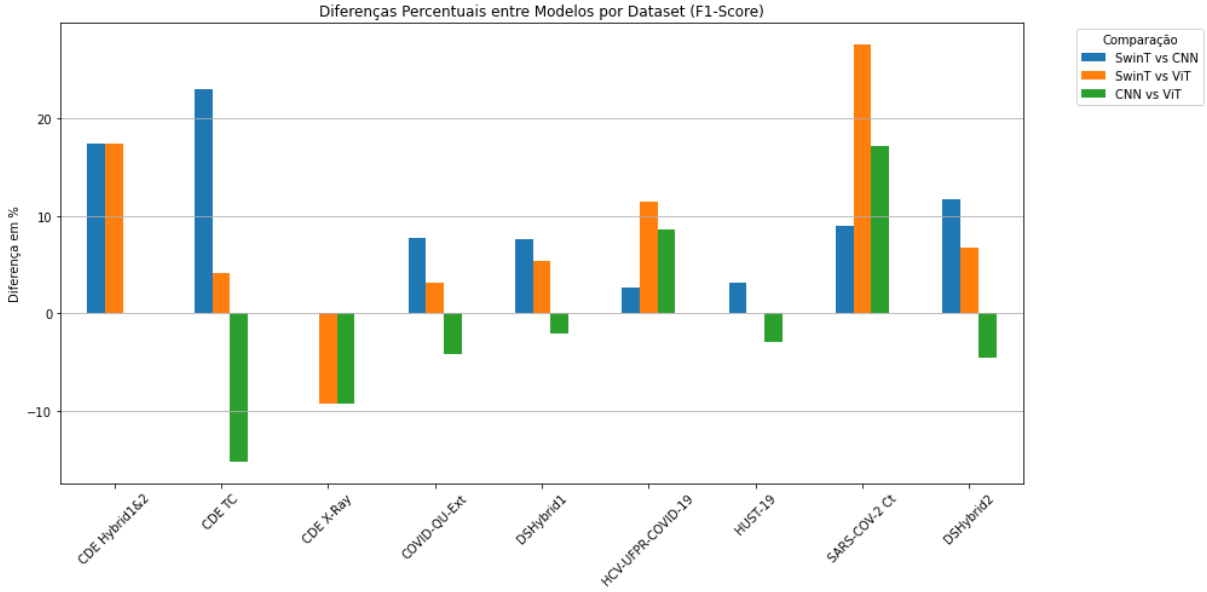
Gráfico 29 – Valores de F1-Score por modelo e *dataset*



Fonte: o autor (2023).

Para complementar a análise, o Gráfico 30 traz a diferença percentual dos valores obtidos para o F1-Score em cada modelo e *dataset*.

Gráfico 30 – Diferenças percentuais por *dataset* para a F1-Score



Fonte: o autor (2023).

Com base nos dados trazidos pelos Gráficos 28, 29 e 30, exploramos e analisamos o desempenho dos três modelos nos nove conjuntos de dados com o objetivo de determinar como esses modelos se saíram em diferentes contextos e identificar suas vantagens e desvantagens em relação ao equilíbrio entre

precisão e sensibilidade em diferentes contextos (F1-Score). Os resultados estão compilados dessa forma:

a) CDE Hybrid1&2:

   1. *SwinT vs*. CNN: *SwinT* têm desempenho superior ao CNN em termos de F1-Score neste conjunto de dados, com pontuações 17.46 pontos percentuais a mais.

   2. *SwinT vs*. ViT: *SwinT* também têm pontuações F1-Score superior ao ViT em 17.46 pontos percentuais neste conjunto.

   3. CNN *vs*. ViT: CNN e ViT obtiveram os mesmos valores de F1-Score. Isso indica que a decisão de escolher um ou outro deve ser tomada avaliando as demais métricas.

b) CDE TC:

   1. *SwinT vs*. CNN: *SwinT* supera o CNN com uma margem de 22.95 pontos em termos de F1-Score. Isso indica que o *SwinT* alcança um melhor equilíbrio entre precisão e sensibilidade neste conjunto de dados.

   2. *SwinT vs*. ViT: *SwinT* supera o ViT com uma margem de 4.17 pontos em termos de F1-Score. Isso sugere que o *SwinT* é a melhor escolha em comparação com o ViT para maximizar o F1-Score.

   3. CNN *vs*. ViT: o CNN é superado pelo ViT com uma margem de 15.28 pontos em termos de F1-Score. Isso mostra que o ViT é significativamente mais competitivo em relação ao CNN neste contexto, mas ainda não alcança um equilíbrio entre precisão e sensibilidade comparável ao do SwinT.

c) CDE *X-Ray*, COVID-QU-Ext, DSHybrid1, HCV-UFPR-COVID-19, HUST-19: para esses conjuntos de dados, os resultados variam, com algumas métricas indicando vantagem para o *SwinT*, ViT ou CNN, dependendo do contexto. Cabe destacar que as diferenças entre eles são menos expressivas, o que não quer dizer que sejam insignificantes.

d) SARS-COV-2 Ct, DSHybrid2: nestes conjuntos de dados, o *SwinT* supera o CNN e o ViT em termos de F1-Score, com margens variadas.

O *SwinT* frequentemente demonstra um equilíbrio promissor entre precisão e sensibilidade em termos de F1-Score, mas existem casos em que o CNN ou o ViT podem ser mais adequados, dependendo do contexto da tarefa.

4.5 DISCUSSÃO

Para facilitar a discussão, inicialmente, devemos responder uma pergunta bastante relevante: Qual é a métrica mais importante no contexto da classificação de imagens radiográfica para auxiliar no diagnóstico de COVID-19?

No contexto de classificação de imagens radiográficas para auxiliar no diagnóstico de COVID-19, é crucial que o modelo tenha um considerável nível de precisão e sensibilidade, especialmente considerando as implicações clínicas e de saúde pública. No entanto, cada métrica tem sua importância:

1. AUC: esta é uma métrica geral que avalia a capacidade do modelo de distinguir entre classes. Um valor de AUC próximo a 1 (um) indica que o modelo tem uma ótima distinção entre as classes, enquanto um valor próximo de 0,5 sugere que o modelo não tem capacidade de distinção melhor do que um chute aleatório. No contexto da COVID-19, um AUC alto é desejável, pois indica que o modelo pode diferenciar bem entre imagens radiográficas de pacientes com e sem COVID-19.
2. acurácia: representa a proporção total de previsões corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de previsões feitas. No entanto, a acurácia pode ser enganosa em conjuntos de dados desequilibrados. Cabe ressaltar que a grande maioria dos conjuntos de dados com imagens médicas contém classes de dados desiquilibrados.
3. precisão: indica a proporção de identificações positivas (previsões de que um paciente tem COVID-19) que estavam realmente corretas. Uma precisão baixa pode levar a muitos falsos positivos, o que, neste contexto, significaria diagnosticar indevidamente muitos pacientes como tendo COVID-19, quando na realidade eles não têm.
4. sensibilidade (*recall*): essa métrica indica a proporção de verdadeiros positivos detectados pelo modelo. No contexto de uma doença potencialmente mortal como a COVID-19, a sensibilidade é extremamente importante, pois um baixo recall significaria que muitos casos verdadeiros de

COVID-19 seriam perdidos (falsos negativos), colocando pacientes e outras pessoas em risco de contágio ao entrarem em contato com esses pacientes que foram erroneamente diagnosticados como estando saudáveis, e foram liberados de tratamento e isolamento. Situação como essa causam aumento do contágio comunitário.

5. F1-Score: é a média harmônica entre precisão e sensibilidade e pode ser uma métrica útil quando há um desequilíbrio nas classes ou quando ambas, precisão e sensibilidade, são importantes.

Dessa forma, para o diagnóstico de COVID-19 usando imagens radiográficas a sensibilidade é, muitas vezes, considerada a métrica mais crítica porque a não detecção de um paciente com COVID-19 (um falso negativo) pode ter consequências graves, como a propagação da doença. Por outro lado, embora falsos positivos (pessoas sem COVID-19 diagnosticadas como positivas) possam levar a estresse e tratamentos desnecessários, do ponto de vista de contenção de uma pandemia, é menos perigoso do que perder um caso verdadeiro.

No entanto, todas as métricas devem ser consideradas em conjunto, e nenhuma métrica por si só pode fornecer uma imagem completa do desempenho de um modelo. O ideal é ter um modelo com alta precisão, alta sensibilidade e um alto valor de AUC.

Para facilitar a discussão e a análise dos resultados, serão apontadas quatro descobertas:

- Descoberta 1: apontada pelos resultados obtidos pelos três modelos para os *datasets* originais (COVID-QU-Ext, HCV-UFPR-COVID-19, HUST-19 e SARS-COV-2 Ct);
- Descoberta 2: apontada pelos resultados obtidos pelos três modelos para os *datasets* híbridos compostos por imagens de Raio X e TC (DSHybrid1 e DSHybrid2);
- Descoberta 3: apontada pelos resultados obtidos pelos três modelos em CDE compostos somente por imagens de Raio X ou somente por imagens de TC ou, ainda, híbrido (CDE Raio X, CDE TC e CDE Hybrid1&2);
- Descoberta 4: apontada pelos resultados obtidos pelos três modelos de forma global, avaliando se é possível apontar um modelo que tenha se saído melhor que os demais avaliados.

4.5.1 Descoberta 1

Após analisar os resultados, podemos observar que todos os modelos (CNN, ViT e *Swin Transformer*) apresentaram um desempenho geralmente alto nos *datasets* originais. No entanto, é importante notar que o desempenho dos modelos pode variar dependendo do *dataset* e do tipo de imagem utilizado.

Analisando os resultados, podemos observar que o modelo *Swin Transformer* obteve um desempenho consistente e superior nos *datasets* originais em relação a acurácia, precisão, sensibilidade e F1-Score. Isso sugere que o *Swin Transformer* é capaz de realizar uma classificação precisa e eficaz entre os casos positivos de COVID-19 e os casos normais nos diferentes tipos de imagem presentes nos *datasets*. Em relação a AUC o *Swin Transformer* empatou com o modelo CNN, mas superou o ViT.

O modelo CNN e o modelo ViT também obtiveram resultados razoavelmente bons nos datasets originais, porém, em comparação com o *Swin Transformer*, apresentaram um desempenho ligeiramente inferior em todas as métricas, exceto na AUC como já destacado. Portanto, em termos de desempenho geral nos *datasets* originais, o modelo *Swin Transformer* se destaca como a escolha mais promissora para a classificação binária de COVID-19 e NORMAL.

No dataset COVID-QU-Ext, todos os modelos demonstraram um desempenho notável, com AUC variando de 0,93 a 0,99. Esse intervalo de valores indica a capacidade eficaz dos modelos em distinguir entre as classes. Além disso, as métricas de acurácia, precisão, *recall* e F1-Score permaneceram consistentemente altas, oscilando entre 0,89 e 1,00. Isso reflete a habilidade dos modelos em realizar classificações precisas, mantendo um equilíbrio sólido entre precisão e *recall*. De maneira geral, esses resultados indicam que os modelos são altamente eficazes na identificação de casos de COVID-19 neste conjunto de dados.

Em contrapartida, no *dataset* HCV-UFPR-COVID-19, o desempenho dos modelos revelou uma maior variabilidade, com valores de AUC variando entre 0,62 e 0,98. Tal diversidade evidencia que alguns modelos possuem uma capacidade discriminatória consideravelmente superior em relação a outros neste *dataset*. Comparativamente com os outros *datasets* originais esses é o mais desafiador e deve possuir algum viés que dificulta os modelos em abstraírem bem

as suas características. As métricas de acurácia, precisão, *recall* e F1-Score também apresentaram uma amplitude notável, oscilando de 0,64 a 0,94. Isso sugere que alguns modelos alcançaram um desempenho substancialmente melhor em comparação aos demais. Em resumo, enquanto alguns modelos tiveram um desempenho relativamente satisfatório, outros não alcançaram resultados satisfatórios neste conjunto de dados. A escolha do modelo mais adequado dependerá, portanto, dos requisitos particulares da aplicação.

No que diz respeito ao *dataset* "HUST-19", todos os modelos exibiram um desempenho excepcional, obtendo valores máximos de AUC, acurácia, precisão, *recall* e F1-Score, todos iguais a 1,00. Comparativamente com os outros *datasets* originais esses é o que teve as melhores métricas sendo alcançadas pelos modelos e deve possuir algum viés que facilita que os modelos abstraíam bem as suas características. Esses resultados indicam que esses modelos são altamente precisos e capazes de realizar classificações perfeitas nesse conjunto de dados. Essa excelência nos resultados sugere que os modelos são altamente eficazes na detecção de casos de COVID-19 neste *dataset* específico.

Por fim, no dataset "SARS-COV-2 Ct", os resultados apresentaram variações entre os modelos, com valores de AUC variando de 0,73 a 0,99. Isso evidencia que alguns modelos demonstraram uma habilidade superior em distinguir valores de Ct do SARS-CoV-2. As métricas de acurácia, precisão, *recall* e F1-Score também revelaram diferenças consideráveis entre os modelos, com algumas variações significativas. De modo geral, os resultados apontam que alguns modelos são mais eficazes nessa métrica, enquanto outros podem não ser tão eficientes.

Resumindo, a escolha do modelo mais apropriado depende dos requisitos específicos da aplicação e dos compromissos entre precisão, *recall* e AUC. Para tarefas onde a máxima precisão é essencial, como na detecção de casos positivos de COVID-19, os modelos ajustados no *dataset* "HUST-19" parecem ser a escolha ideal, uma vez que apresentaram resultados perfeitos em todas as métricas. Entretanto, é fundamental levar em consideração o contexto e os objetivos da aplicação ao optar pelo modelo mais adequado.

4.5.2 Descoberta 2

Observando os resultados, pode-se notar que os três modelos obtiveram um desempenho geralmente alto nos *datasets* híbridos. No entanto, em termos de acurácia, precisão, sensibilidade e F1-Score, o modelo *Swin Transformer* obteve resultados consistentemente melhores em comparação com os outros dois modelos.

Esses resultados sugerem que o modelo *Swin Transformer* possui uma capacidade robusta de classificação nos *datasets* híbridos, indicando sua eficácia na tarefa de classificação binária de COVID-19 e NORMAL nos *datasets* híbridos. O *Swin Transformer* demonstrou um desempenho superior em todas as métricas avaliadas, incluindo AUC, acurácia, precisão, sensibilidade e F1-Score, o que sugere uma melhor capacidade de discriminar corretamente entre os casos positivos de COVID-19 e os casos normais.

Analisando mais detalhadamente cada *dataset* é possível verificar que os resultados para o DSHybrid1 revelam um desempenho notável em todas as métricas avaliadas. Todos os modelos demonstraram uma capacidade excepcional de distinguir efetivamente entre as classes, conforme evidenciado pelos valores de AUC variando de 0,93 a 0,99. Além disso, as métricas de acurácia, precisão, *recall* e F1-Score também exibiram resultados promissores, com valores próximos a 0,99. Esses resultados coletivos refletem a alta precisão e capacidade de classificação dos modelos nesse *dataset*, sugerindo que eles são altamente eficazes na detecção das classes relevantes para a aplicação.

No *dataset* DSHybrid2, os modelos apresentaram um desempenho sólido, embora ligeiramente inferior em comparação com o DSHybrid1. Os valores de AUC variaram de 0,84 a 0,97, indicando que os modelos ainda têm uma capacidade promissora de discriminação entre as classes. As métricas de acurácia, precisão, *recall* e F1-Score também mantiveram um nível elevado, com valores variando de 0,84 a 0,97. Isso sugere que os modelos são precisos na classificação das amostras nesse *dataset*, apesar de uma pequena diferença em relação ao DSHybrid1.

Em resumo, tanto o DSHybrid1 quanto o DSHybrid2 apresentaram resultados de métricas positivos, com modelos que demonstram uma capacidade sólida de classificação. Embora o DSHybrid1 tenha obtido resultados ligeiramente

melhores, o DSHybrid2 ainda se destaca como uma escolha eficaz para treinar os modelos para tarefas de classificação.

4.5.3 Descoberta 3

Ao analisar com mais profundidade os resultados obtidos pelos três modelos nas cinco métricas nesses três *datasets* (CDE TC, CDE Raio X e CDE Hybrid1&2) é possível verificar que os desempenhos de todos eles são significativamente piores que os obtidos nos *datasets* originais e híbridos. Isso demonstra que a estratégia de CDE certamente é um desafio significativo para qualquer um dos três modelos.

Ao analisar esses resultados desse cenário de experimento, é importante considerar as métricas relevantes para a aplicação específica e os requisitos clínicos. No geral, o modelo *SwinT* obteve um desempenho ligeiramente melhor em termos de acurácia, precisão, sensibilidade e F1-Score nos três *datasets* avaliados em CDE em comparação com os modelos CNN e ViT. Quanto à AUC os três modelos se equiparam.

No contexto de imagens médicas, especialmente na detecção de COVID-19 a partir de imagens radiográficas, a escolha do modelo de DL e sua eficácia pode ter implicações significativas. Estas imagens são críticas, pois os diagnósticos precisos podem acelerar o tratamento dos pacientes, isolar casos positivos para evitar a propagação e, eventualmente, salvar vidas.

Ao comparar os três modelos — *SwinT*, CNN e ViT — observa-se diferenças estatisticamente significativas nas métricas de desempenho entre o SwinT e CNN, com exceção da AUC. Quando se comparam os três, o *SwinT* leva a melhor em todas as métricas.

No entanto, para se determinar qual é o potencial melhor modelo a ser escolhido é importante destacar algumas considerações fundamentais:

1. contexto da aplicação: em cenários médicos como este, a precisão é crucial. Mesmo uma melhoria aparentemente pequena de 4,94% na AUC entre ViT e os outros dois modelos podem ser de grande valor clínico. Uma melhoria dessa magnitude pode resultar em um número significativo de casos detectados corretamente, levando a intervenções médicas oportunas.

2. base de comparação: notavelmente, enquanto *SwinT* e CNN mostram igualdade na métrica AUC, *SwinT* supera a CNN em mais de 10% nas métricas de Acurácia, Precisão, Sensibilidade e F1-Score, além de apresentar diferenças estatisticamente significativas. Essa é uma diferença considerável, particularmente considerando que ambos começam de uma base já alta, indicando a robustez e a superioridade potencial do modelo *SwinT* para esta tarefa.
3. *trade-offs*: embora não detalhados aqui, é crucial considerar quaisquer *trade-offs* associados ao uso do modelo *SwinT*. Por exemplo, os estudos apontaram que ele exige mais recursos computacionais e tempo para treinar ou inferir do que a CNN ou o ViT. No ambiente clínico, o tempo pode ser essencial, e qualquer atraso no diagnóstico pode ter implicações graves.
4. consistência entre métricas: os resultados apontaram que melhoria de *SwinT* sobre a CNN é consistente em todas as métricas, exceto na AUC onde há um empate, o que sugere uma melhoria geral em sua capacidade de detectar a COVID-19 em imagens radiográficas.
5. estabilidade da melhoria: muito embora o estudo tenha utilizado quatro bases de dados originais que foram combinadas de diferentes formas e resultaram em nove *datasets* distintos, seria importante expandir a avaliação da robustez dessas métricas em diferentes conjuntos de dados ou para garantir que essas melhorias não sejam anomalias ou consequências de algum viés nos conjuntos de dados.
6. custo da implementação: dado o cenário crítico de detecção da COVID-19, pode-se argumentar que qualquer melhoria, independentemente do custo, vale a pena. No entanto, é essencial ser pragmático. Se *SwinT* requer uma infraestrutura significativamente mais sofisticada ou mais dados para treinamento, os centros médicos com recursos limitados podem achar difícil adotá-lo. A título de exemplo, observamos que para executar o treinamento no maior *dataset* utilizado no estudo (CDE_Hybrid1&2 contendo 39.631 imagens) o tempo gasto por cada modelo foi o seguinte: *SwinT* — 3 horas, 53 minutos e 32 segundos; ViT — 3 horas, 4 minutos e 28.63 segundos; enquanto o CNN demorou 1 hora, 47 minutos e 6.96 segundos.

Em resumo, enquanto *SwinT* demonstra ser um modelo promissor com suas métricas aprimoradas, é essencial ponderar estas melhorias contra quaisquer *trade-offs* associados. Os resultados da análise por *dataset*, por exemplo, enfatizaram a necessidade de escolher o modelo de rede neural conforme o conjunto de dados e a métrica desejada. Enquanto o *SwinT* demonstrou nos testes estatísticos que as métricas de acurácia, precisão, sensibilidade e F1-Score, possuem diferenças estatísticas significativa (melhor) em várias estratégias de execução, o CNN ou ViT podem ser mais apropriados em casos específicos.

Dada a gravidade da detecção da COVID-19 e o impacto potencial na saúde pública, estas descobertas sugerem uma exploração ainda mais aprofundada do modelo *SwinT*, para garantir a eficácia e viabilidade em ambientes clínicos reais.

Abaixo seguem observações específicas sobre cada *dataset* que empregou a técnica CDE:

Dataset CDE Raio X

Os resultados para o *dataset* CDE Raio X revelaram uma variabilidade no desempenho dos modelos. A métrica AUC varia de 0,46 a 0,55, sugerindo que os modelos têm um poder discriminatório limitado na diferenciação das classes neste conjunto de dados. Isso é consistente com os desafios comuns enfrentados na análise de imagens médicas, onde a complexidade da tarefa pode afetar o desempenho. De longe esse é o *dataset* onde foi obtido os piores resultados por todos os modelos em todas as métricas e em vários casos os valores obtidos nas métricas (menores que 0,5) são piores que os obtidos pelo lançamento de uma moeda para tirar cara ou coroa, onde a probabilidade é de 0,5 para cara e 0,5 para coroa.

As métricas de acurácia, precisão, *recall* e F1-Score também variam, mas em geral, os valores estão na faixa de 0,27 a 0,71. Isso indica que alguns modelos conseguem classificar amostras com uma precisão razoável, enquanto outros têm um desempenho bastante inferior. No geral, o desempenho nesse dataset é o mais desafiador, e a escolha do modelo dependerá dos requisitos específicos da aplicação e das tolerâncias a falsos positivos e falsos negativos.

Dataset CDE TC

O *dataset* CDE TC mostra resultados variados, com AUC variando de 0,28 a 0,79. Isso sugere que alguns modelos têm uma capacidade melhor de discriminar entre as classes neste conjunto de dados em comparação com outros. O desafio apresentado por esse *dataset* é evidente na variabilidade das métricas de acurácia, precisão, *recall* e F1-Score, que variam de 0,57 a 0,79. Alguns modelos alcançam um desempenho razoável, enquanto outros têm resultados menos satisfatórios. Para os casos em que as métricas são inferiores a 0,5 vale a mesma observação realizada para o dataset CDE Raio X.

Em resumo, o *dataset* CDE TC se apresenta como desafiador para a tarefa de classificação, e a escolha do modelo dependerá da importância relativa de diferentes métricas de desempenho, como acurácia, precisão, ou sensibilidade, para a aplicação específica.

Dataset CDE Hybrid1&2

O *dataset* CDE Hybrid1&2 exibe resultados mistos. A métrica AUC varia de 0,63 a 0,76, sugerindo uma capacidade moderada de discriminação entre as classes. As métricas de acurácia, precisão, *recall* e F1-Score também mostram variação, com valores de 0,61 a 0,77.

Isso indica que alguns modelos têm um desempenho razoável na classificação desses dados híbridos, enquanto outros podem precisar de otimização adicional. A escolha do modelo nesse *dataset* dependerá das necessidades específicas da aplicação e dos *trade-offs* entre as métricas de desempenho.

Em resumo, o *dataset* "CDE Hybrid1&2" apresenta desafios na tarefa de classificação, e a seleção do modelo adequado dependerá da prioridade dada a diferentes métricas de desempenho e da necessidade de otimização adicional.

4.5.4 Descoberta 4

Ao analisar de forma global as métricas obtidas pelos modelos CNN, ViT e *Swin Transformer* na classificação binária de COVID-19 e NORMAL para determinar qual modelo teve a melhor performance, podemos fazer as seguintes observações:

4.5.4.1 Considerações sobre os tipos de imagens

Os resultados também destacam a importância do tipo de imagem na classificação. Os modelos tiveram um desempenho geralmente melhor com imagens de TC em comparação com imagens de Raio X. Portanto, pode-se inferir que a inclusão de imagens de TC no conjunto de treinamento pode melhorar o desempenho dos modelos devido ao fato de serem mais ricas em características que podem ser extraídas pelos modelos.

4.5.4.2 Desempenho dos Modelos CNN

O modelo CNN demonstrou resultados sólidos em diversos *datasets* e tipos de imagem. No entanto, embora tenha registrado valores elevados de AUC, ficou aquém nas métricas de acurácia, precisão, sensibilidade e F1-Score, principalmente quando comparado ao *SwinT* na classificação entre COVID-19 e NORMAL. Maiores informações sobre esse ponto estão no subcapítulo 4.4.2.2.

É válido salientar que o HCV-UFPR-COVID-19 foi o *dataset* mais complexo para todos os modelos, visto que ele exibiu métricas inferiores em relação aos outros três datasets originais, sendo eles: COVID-QU-Ext, Hust-19 e SARS-COV2-Ct. Suspeitamos que esse conjunto de dados exerça uma influência considerável nos resultados dos modelos em outros *datasets* que usaram dados do HCV-UFPR-COVID-19, como o CDE *X-Ray* (que apresentou os piores resultados em todas as métricas e modelos do estudo) e o DSHybrid2 que sofreu degradação, porém não tão expressiva. Em contrapartida, o DSHybrid1, que mesclou dados do COVID-QU-Ext (Raio X) e do HUST-19 (TC), apresentou desempenho superior ao DSHybrid2. Tal fato sugere que os conjuntos de dados podem carregar vieses que interferem nos resultados quando mesclados.

Contudo, a formação dos *datasets* híbridos aparenta ter atenuado vieses mais acentuados, como os observados no HCV-UFPR-COVID-19, resultando em desempenhos até melhores que os dos *datasets* originais mais desafiadores, quando avaliados individualmente. Um aspecto importante a considerar é a mescla de imagens de TC e raio X nos *datasets* híbridos. Notou-se que os modelos costumam ser mais precisos ao trabalhar com imagens de TC em comparação com as de raio X. Quando essas imagens são combinadas, o desempe-

nho tende a superar o uso exclusivo de Raio X, mas ainda assim fica aquém se comparado ao *dataset* composto somente por imagens de TC.

4.5.4.3 Desempenho dos Modelos ViT

O modelo ViT também apresentou resultados promissores, com valores elevados de AUC, acurácia, precisão, sensibilidade e F1-Score. Eles demonstraram boa capacidade de classificação tanto em imagens de Raio X quanto em imagens de TC. No entanto, assim como os modelos CNN, o desempenho foi significativamente inferior no *dataset* HCV-UFPR-COVID-19, quando comparado com os demais, consequentemente também sofreu importante degradação no CDE Raio X.

4.5.4.4 Desempenho dos Modelos *Swin Transformer*

O modelo *Swin Transformer* mostrou um desempenho consistente e forte na classificação binária de COVID-19 e NORMAL. Ele alcançou valores altos em todas as métricas avaliadas em diferentes *datasets* e tipos de imagem. O *Swin Transformer* teve um desempenho particularmente robusto no *dataset* HUST-19, obtendo valores próximos de um (1) para todas as métricas.

4.5.4.5 Comparação entre os Modelos

Comparando os modelos CNN, ViT e *Swin Transformer*, observamos que todos tiveram um desempenho promissor na classificação binária. No entanto, o *Swin Transformer* mostrou um desempenho mais consistente e superior em vários *datasets* e tipos de imagem, especialmente no *dataset* HUST-19. Isso sugere que o *Swin Transformer* pode ser uma escolha promissora para a classificação precisa de COVID-19 e NORMAL. Além disso, o *SwinT* destaca-se indiscutivelmente na métrica considerada a mais crucial para a análise de um modelo de classificação de imagens médicas: a sensibilidade.

A sensibilidade é uma métrica crucial na classificação de COVID-19, pois indica a capacidade do modelo de identificar corretamente os casos positivos da doença. Um modelo com alta sensibilidade é capaz de minimizar os casos de falsos negativos, ou seja, reduzir a quantidade de casos de COVID-19 que são erroneamente classificados como NORMAL. Portanto, modelos com alta sensibi-

lidade são preferíveis em aplicações de detecção de COVID-19. Da compilação de dados pode-se observar que os modelos CNN, ViT e *Swin Transformer* apresentaram um potencial promissor para a classificação binária de COVID-19 e NORMAL. No entanto, o *Swin Transformer* se destacou com um desempenho geralmente superior em termos de sensibilidade. Maiores detalhes podem ser obtidos no subcapitulo 4.4.2.2.

Finalmente, é importante destacar que os resultados experimentais indicaram um desempenho mais elevado nas métricas do modelo *Swin Transformer*. Isso está alinhado com as observações reportadas na literatura, conforme corroborado por autores como Liu *et al*. (2021) e Dosovitskiy *et al*. (2021) no contexto da detecção e diagnóstico de COVID-19, sobretudo em *datasets* híbridos.

Uma das possíveis causas para o desempenho superior alcançado pelo *SwinT* pode ser devido à sua capacidade de aprender representações complexas e identificar padrões sutis. O *Swin Transformer* foi projetado para capturar relações de longo alcance entre diferentes partes da imagem, permitindo uma compreensão mais abrangente das informações contextuais, conforme defendido por Vaswani *et al*. (2017).

Embora o *SwinT* tenha apresentado valores elevados em todas as métricas nos experimentos, a escolha do modelo ideal para tarefas em ambientes médicos reais deve considerar fatores como disponibilidade de dados, recursos computacionais e requisitos clínicos. Portanto, validações adicionais e comparações de desempenho em diferentes cenários médicos de utilização real são necessárias para validar a capacidade de generalização e aplicação prática.

Ainda no contexto da discussão cabe ressaltar que os conjuntos de dados desempenham um papel crucial no treinamento e validação de modelos de *machine/deep learning* para a classificação de imagens médicas. Os *datasets* utilizados neste estudo — COVID-QU-Ext, HCV-UFPR-COVID-19, HUST-19, SARS-COV-2 Ct, DSHybrid1, DSHybrid2, CDE TC, CDE Raio X e CDE Hybrid1&2 — abrangem uma diversidade de imagens, incluindo radiografias e TC, o que permite uma visão abrangente e variada do quadro clínico de COVID-19. Cada um desses conjuntos de dados possui suas próprias particularidades e vieses, como a diversidade de pacientes, as condições sob as quais as imagens

foram obtidas e o tipo de imagem (Raio X ou TC), o que pode ter impactos significativos nos resultados alcançados pelos modelos.

O viés no treinamento dos modelos pode surgir de várias formas, por exemplo, se um conjunto de dados tiver uma grande proporção de imagens de pacientes com casos graves de COVID-19 em comparação com casos leves ou normais, isso pode levar o modelo a ter um desempenho superior na identificação de casos graves, mas inferior na identificação de casos leves ou normais. Portanto, é crucial garantir que os *datasets* sejam equilibrados e representem uma variedade de cenários clínicos, o que não foi possível garantir no presente trabalho tendo em vista que os conjunto de dados já estavam formados e nos ativemos a utilizá-lo nos experimentos.

No que diz respeito ao tipo de imagem, as imagens de Raio X e as imagens de TC proporcionam diferentes níveis de detalhes e informações diagnósticas. As imagens de Raio X são mais simples e acessíveis, mas podem não revelar tantos detalhes quanto uma imagem de TC, que oferece uma visão tridimensional e de alta resolução do tórax. Neste estudo, foi observado que os modelos geralmente tiveram um desempenho melhor com imagens de TC em comparação com as imagens de Raio X. Isso sugere que as imagens de TC podem conter mais características úteis e discriminativas que ajudam os modelos a fazerem uma distinção mais precisa entre casos COVID-19 e normais.

Além disso, os *datasets* híbridos, que contêm tanto imagens de Raio X quanto de TC, oferecem uma oportunidade para treinar modelos que podem funcionar eficazmente em diferentes tipos de imagem. No entanto, vale a pena notar que a mistura de diferentes tipos de imagens em um único conjunto de dados pode introduzir complexidade adicional e desafios para o treinamento do modelo.

Ao analisar as três distintas abordagens de mescla de *datasets*, sendo a primeira voltada aos datasets originais, a segunda focada nos *datasets* híbridos e a terceira denominada de CDE, identificamos diferentes padrões. Nos datasets originais, as métricas mostraram-se satisfatórias. Nos datasets híbridos, na maior parte dos casos, as métricas se equipararam ou até superaram as dos datasets originais. Contudo, ao avaliarmos a estratégia de CDE, percebemos uma degradação notável em relação às duas primeiras abordagens. Isso indica

que essa última estratégia apresenta desafios significativos, dificultando o aprendizado pelos modelos adotados neste estudo.

As implicações desses fatores para a classificação de imagens médicas são significativas. Ao desenvolver e avaliar modelos de classificação, é essencial considerar a diversidade e representatividade dos conjuntos de dados, os potenciais vieses introduzidos pelas características dos *datasets*, e a influência do tipo de imagem sobre o desempenho do modelo. Ao levar em conta essas questões, é possível desenvolver modelos mais robustos e precisos para a classificação de imagens médicas.

De acordo com nossa análise, o *Swin Transformer* destaca-se globalmente como o modelo mais promissor dentre os três analisados. Sua mediana superior nas métricas e a baixa variabilidade demonstram sua robustez e eficácia em diferentes cenários. Este fato está alinhado com o que se observa no estado da arte atual da visão computacional, onde o *SwinT* vem recebendo atenção crescente devido ao seu desempenho superior.

Entretanto, é fundamental reconhecer que a eficácia de um modelo pode ser influenciada por fatores específicos do *dataset* ou da aplicação em questão. Portanto, é crucial avaliar modelos com base no contexto em que serão implementados, bem como permanecer atento às evoluções e inovações no campo da visão computacional e aprendizado de máquina.

Em conclusão, ao compararmos os resultados de nossos experimentos com o estado da arte, percebemos que são promissores e competitivos. Contudo, é essencial salientar que diversos fatores, como as condições experimentais, as versões dos modelos e os ajustes de hiperparâmetros, podem afetar os resultados. Uma análise mais aprofundada exigiria detalhes adicionais sobre a execução de cada experimento.

Revisitando a seção 1.3 (Resultados Obtidos), e tendo em vista o panorama apresentado na revisão da literatura e os insights gerados pelos experimentos, concluímos que a utilização do *Swin Transformer* tem o potencial de se tornar uma ferramenta valiosa no auxílio ao aprimoramento do diagnóstico. Isso pode se refletir em diversos aspectos, desde auxílio na melhoria do diagnóstico, redução de erros médicos, aceleração do processo e redução de custos no processo de diagnósticos da COVID-19.

No que tange à contribuição ao universo acadêmico e para a pesquisa científica, este trabalho se destaca não apenas pelos resultados inéditos alcançados, mas também pelo seu caráter colaborativo. Os achados deste estudo têm o potencial de enriquecer a base de conhecimentos da comunidade científica, proporcionando *insights* valiosos para o refinamento de técnicas e algoritmos já existentes. Além disso, esses resultados abrem caminho para sua aplicação prática em ensaios clínicos e em contextos médicos reais, ampliando ainda mais o escopo e o impacto do presente estudo no campo da visão computacional aplicada ao domínio das imagens médicas.

5 CONCLUSÃO E TRABALHOS FUTUROS

Este estudo examinou a aplicação de três modelos de aprendizado de profundo na detecção e diagnóstico da COVID-19, foram eles *Resnet50*, ViT e *Swin Transformer*, tendo o último se destacado pelo desempenho. Este modelo mostrou eficácia particularmente notável nos conjuntos de dados originais, híbridos e CDE.

As possíveis razões para as diferenças observadas nas métricas de desempenho entre os modelos *Swin Transformer* (*SwinT*), redes neurais convolucionais (CNN) e *Vision Transformer* (ViT) no contexto da classificação de imagens médicas para a detecção da COVID-19 podem ser as seguintes:

a) os *Swin Transformers*, com sua arquitetura centrada em mecanismos de atenção, têm a capacidade distinta de identificar relações de longo alcance em imagens, uma característica crucial ao analisar imagens médicas onde detalhes cruciais podem ser sutis.

b) adicionalmente, sua habilidade de discernir representações hierárquicas em múltiplas escalas torna-os aptos a reconhecer nuances, como texturas e detalhes microscópicos.

c) esta arquitetura também exibe uma notável capacidade de generalização, uma vantagem em cenários médicos com *datasets* limitados.

d) contudo, quando observamos a métrica AUC, uma medida que avalia a precisão do modelo em diferentes limiares de probabilidade, a superioridade dos *Transformers* não é tão marcante. Esta discrepância pode ser atribuída ao modo como o Transformer calibra probabilidades.

A arquitetura do *Swin Transformer*, fundamentada em *Transformers*, em hipótese, facilitou a detecção de padrões sutis e características relevantes nas imagens, independente da modalidade de *dataset* analisado ou tipo de imagem utilizada. O modelo demonstrou notável sensibilidade, indicando uma taxa de falsos negativos mais baixa na classificação da COVID-19, elemento crucial para um diagnóstico adequado.

Portanto, ao decidir entre *Swin Transformers*, CNN e ViT, é essencial considerar tanto as métricas de desempenho quanto as necessidades clínicas específicas.

A seleção do modelo de DL mais apropriado para uma tarefa clínica específica é uma decisão multifacetada que não deve ser tomada com base apenas no desempenho bruto. Fatores como a quantidade e qualidade dos dados disponíveis, a capacidade de processamento e os objetivos clínicos precisam ser meticulosamente avaliados. O *Swin Transformer*, embora tenha demonstrado um desempenho impressionante em algumas métricas, ainda requer validações adicionais em cenários variados. As métricas não devem ser vistas isoladamente, mas sim em conjunto, para entender a eficácia global de um modelo. Ademais, é vital comparar a eficácia dos diferentes modelos em ambientes clínicos diversificados para garantir que sejam robustos e versáteis.

Em suma, o *Swin Transformer*, apesar de suas métricas notáveis, não é necessariamente o padrão-ouro para todas as aplicações clínicas. A seleção da rede neural ideal deve ser fundamentada no *dataset* em questão e na métrica que é prioritária para otimização. Mesmo com o *SwinT* se destacando em métricas como AUC, precisão e sensibilidade, existem contextos nos quais outras arquiteturas, como CNN ou ViT, podem se mostrar mais propícias. Dessa forma, é imperativo que pesquisadores e profissionais da saúde avaliem profundamente o cenário clínico e as demandas específicas antes de se decidirem por uma abordagem de modelagem particular.

No âmbito de trabalhos futuros, ressalta-se a importância de ensaios clínicos desses modelos em ambientes reais como clínicas de radiologia e hospitais. Ademais, seria benéfico dispor de grandes conjuntos de dados, compostos por milhões de imagens médicas de diferentes origens, para acelerar a aplicação dessas tecnologias na área médica. Nesse sentido, durante o trabalho houve tratativas com uma clínica de radiologia de renome em Curitiba. Porém, a clínica abandou o diálogo sobre o projeto, mesmo sem custo para ela, e não respondeu as frequentes tentativas de dar andamento no projeto. Isso demonstra que ainda existem significativas resistências culturais e barreiras de entrada para a efetiva aplicação desse tipo de pesquisa inovadora no ambiente médico real.

Por fim, é imprescindível expandir a pesquisa para um maior número de cenários e modelos, fornecendo uma compreensão mais abrangente das relações investigadas neste trabalho. A análise da aplicação prática desses modelos no ambiente clínico configura uma importante linha de pesquisa futura. Este

estudo, ao destacar a complexidade do tema, reitera a necessidade de continuar explorando novas abordagens que possam contribuir para o avanço da pesquisa no campo da visão computacional aplicada ao domínio das imagens médicas, beneficiando assim a saúde pública de maneira significativa.

Uma abordagem possível e interessante pode ser comparar os resultados obtidos pelos modelos utilizados no presente estudos com os obtidos pelo novo modelo *Swin Transformer* V2, entre outros, bem como empregar técnicas de Inteligência Artificial Explicável (XAI), tais como *Class Activation Map* (Grad-CAM), *Local Interpretable Model-agnostic Explanations* (LIME) e *t-distributed Stochastic Neighbor Embedding* (t-SNE) aos experimentos.

REFERÊNCIAS

ABIYEV, Rahib H.; ISMAIL, Abdullahi. COVID-19 and Pneumonia Diagnosis in X-Ray Images Using Convolutional Neural Networks. **Mathematical Problems in Engineering**, [S.L.], v. 2021, p. 1-14, 24 nov. 2021. Hindawi Limited. http://dx.doi.org/10.1155/2021/3281135.

AGGARWAL, Priya; MISHRA, Narendra Kumar; FATIMAH, Binish; SINGH, Pushpendra; GUPTA, Anubha; JOSHI, Shiv Dutt. COVID-19 image classification using deep learning: advances, challenges and opportunities. **Computers in Biology and Medicine**, [S.L.], v. 144, p. 105350, maio 2022. Elsevier BV. http://dx.doi.org/10.1016/j.compbiomed.2022.105350.

AHAMED, Khabir Uddin; ISLAM, Manowarul; UDDIN, Ashraf; AKHTER, Arnisha; PAUL, Bikash Kumar; YOUSUF, Mohammad Abu; UDDIN, Shahadat; QUINN, Julian M.W.; MONI, Mohammad Ali. A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images. **Computers in Biology and Medicine**, [S.L.], v. 139, p. 105014, dez. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.compbiomed.2021.105014.

AKIBA, Takuya; SANO, S; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. Optuna: A Next-generation Hyperparameter Optimization Framework**.** *In:* Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining**, Anais...,** [S.l.], 2019. p. 2623-2631. ACM. DOI: 10.1145/3292500.3330701.

ARIAS-GARZÓN, Daniel; TABARES-SOTO, Reinel; BERNAL-SALCEDO, Joshua; RUZ, Gonzalo A.. Biases associated with database structure for COVID-19 detection in X-ray images. **Scientific Reports**, [S.L.], v. 13, 1 mar. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-023-30174-1.

ASIF, Sohaib; ZHAO, Ming; TANG, Fengxiao; ZHU, Yusen. A deep learning-based framework for detecting COVID-19 patients using chest X-rays. **Multimedia Systems**, [S.L.], v. 28, n. 4, p. 1495-1513, 22 mar. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00530-022-00917-7.

AWAN, Mazhar Javed; IMTIAZ, Muhammad Wasif; USAMA, Muhammad; REHMAN, Amjad; AYESHA, Noor; SHEHZAD, Hafiz Muhammad Faisal. Covid-19 Detection by using Deep learning-based Custom Convolution Neural Network (CNN). 2021 International Conference on Innovative Computing (ICIC), **Anais...**, [S.L.], 9 nov. 2021. IEEE. http://dx.doi.org/10.1109/icic53490.2021.9693071.

BAKATOR, Mihalj; RADOSAV, Dragica. Deep Learning and Medical Diagnosis: a review of literature. **Multimodal Technologies and Interaction**, [S.L.], v. 2, n. 3, p. 47, 17 ago. 2018. MDPI AG. http://dx.doi.org/10.3390/mti2030047.

BANERJEE, Avinandan; SARKAR, Arya; ROY, Sayantan; SINGH, Pawan Kumar; SARKAR, Ram. COVID-19 chest X-ray detection through blending ensem-

ble of CNN snapshots. **Biomedical Signal Processing and Control**, [S.L.], v. 78, n. 1, p. 104000, set. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.bspc.2022.104000.

BASILI, Victor R.; CALDIERA, Gianluigi; ROMBACH, Dieter H. The Goal Question Metric Approach. *In*: MARCINIAK, John J. (Editor-in-Chief). **Encyclopedia of Software Engineering**. Vol 1., 1994.

CAO, Kai; DENG, Tao; ZHANG, Chuanlin; LU, Limeng; LI, Lin. A CNN-transformer fusion network for COVID-19 CXR image classification. **Plos One**, [S.L.], v. 17, n. 10, p. e0276758, 27 out. 2022. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0276758.

CASTIGLIONE, Aniello; VIJAYAKUMAR, Pandi; NAPPI, Michele; SADIQ, Saima; UMER, Muhammad. COVID-19: automatic detection of the novel coronavirus disease from CT images using an optimized convolutional neural network. **IEEE Transactions on Industrial Informatics**, [S.L.], v. 17, n. 9, p. 6480-6488, set. 2021. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tii.2021.3057524.

CASTRO, Rodolfo; LUZ, Paula M.; WAKIMOTO, Mayumi D.; VELOSO, Valdilea G.; GRINSZTEJN, Beatriz; PERAZZO, Hugo. COVID-19: a meta-analysis of diagnostic test accuracy of commercial assays registered in Brazil. **The Brazilian Journal Of Infectious Diseases**, (S.L.), v. 24, n. 2, p. 180-187, mar. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.bjid.2020.04.003.

CATALA, Omar del Tejo; IGUAL, Ismael Salvador; PEREZ-BENITO, Francisco Javier; ESCRIVA, David Millan; CASTELLO, Vicent Ortiz; LLOBET, Rafael; PEREZ-CORTES, Juan-Carlos. Bias Analysis on Public X-Ray Image Datasets of Pneumonia and COVID-19 Patients. **IEEE Access**, [S.L.], v. 9, p. 42370-42383, 2021. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/access.2021.3065456.

CHEN, Leiyu; LI, Shaobo; BAI, Qiang; YANG, Jing; JIANG, Sanlong; MIAO, Yanming. Review of Image Classification Algorithms Based on Convolutional Neural Networks. **Remote Sensing**, [S.L.], v. 13, n. 22, p. 4712, 21 nov. 2021. MDPI AG. http://dx.doi.org/10.3390/rs13224712.

CHEN, Hui; ZHANG, Tian; CHEN, Runbin; ZHU, Zihang; WANG, Xu. A Novel COVID-19 Image Classification Method Based on the Improved Residual Network. **Electronics**, [S.L.], v. 12, n. 1, p. 80, 25 dez. 2022. MDPI AG. http://dx.doi.org/10.3390/electronics12010080.

CHENG, Phillip M.; MONTAGNON, Emmanuel; YAMASHITA, Rikiya; PAN, Ian; CADRIN-CHêNEVERT, Alexandre; ROMERO, Francisco Perdigón; CHARTRAND, Gabriel; KADOURY, Samuel; TANG, An. Deep Learning: an update for radiologists. **Radiographics**, [S.L.], v. 41, n. 5, p. 1427-1445, set. 2021. Radiological Society of North America (RSNA). http://dx.doi.org/10.1148/rg.2021200210.

CHETOUI, Mohamed; AKHLOUFI, Moulay A.. Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays. **Journal of Clinical Medicine**, [S.L.], v. 11, n. 11, p. 30131-1, 26 maio 2022. MDPI AG. http://dx.doi.org/10.3390/jcm11113013.

CHOLLET, Francois. Xception: deep learning with depthwise separable convolutions. **IEEE Conference on Computer Vision and Pattern Recognition** (CVPR), Anais… [S.L.], jul. 2017. IEEE. http://dx.doi.org/10.1109/cvpr.2017.195.

CIRESAN, D.; MEIER, U.; SCHMIDHUBER, J.. Multi-column deep neural networks for image classification. **IEEE Conference on Computer Vision And Pattern Recognition**, Anais… (S.L.), jun. 2012. IEEE. http://dx.doi.org/10.1109/cvpr.2012.6248110.

COSTA, Yandre M. G.; SILVA, Sergio A.; TEIXEIRA, Lucas O.; PEREIRA, Rodolfo M.; BERTOLINI, Diego; BRITTO, Alceu S.; OLIVEIRA, Luiz S.; CAVALCANTI, George D. C.. COVID-19 Detection on Chest X-ray and CT Scan: a review of the top-100 most cited papers. **Sensors**, [S.L.], v. 22, n. 19, p. 7303, 26 set. 2022. MDPI AG. http://dx.doi.org/10.3390/s22197303.

DEHKORDI, Hojat Asgarian; KASHIANI, Hossein; IMANI, Amir Abbas Hamidi; SHOKOUHI, Shahriar Baradaran. Lightweight Local Transformer for COVID-19 Detection Using Chest CT Scans. **11Th International Conference on Computer Engineering and Knowledge (ICCKE)**, Anais… [S.L.], v. 1, n. 1, p. 1-1, 28 out. 2021. IEEE. http://dx.doi.org/10.1109/iccke54056.2021.9721517.

DINH, Tuan Le; LEE, Suk-Hwan; KWON, Seong-Geun; KWON, Ki-Ryong. COVID-19 Chest X-ray Classification and Severity Assessment Using Convolutional and Transformer Neural Networks. **Applied Sciences**, [S.L.], v. 12, n. 10, p. 4861, 11 maio 2022. MDPI AG. http://dx.doi.org/10.3390/app12104861.

DOSOVITKIY, Alexey; *et al*. An image is worth 16x16 words: transformers for image recognition at scale. *In*: **ICLR 2021**, Anais… https://doi.org/10.48550/arXiv.2010.11929.

ESTEVA, Andre; ROBICQUET, Alexandre; RAMSUNDAR, Bharath; KULESHOV, Volodymyr; DEPRISTO, Mark; CHOU, Katherine; CUI, Claire; CORRADO, Greg; THRUN, Sebastian; DEAN, Jeff. A guide to deep learning in healthcare. **Nature Medicine**, [S.L.], v. 25, n. 1, p. 24-29, jan. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41591-018-0316-z.

FENG, Xin; JIANG, Youni; YANG, Xuejiao; DU, Ming; LI, Xin. Computer vision algorithms and hardware implementations: a survey. **Integration**, [S.L.], v. 69, p. 309-320, nov. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.vlsi.2019.07.005.

FREIRE-OBREGÓN, David; MARSICO, Maria de; BARRA, Paola; LORENZO-NAVARRO, Javier; CASTRILLÓN-SANTANA, Modesto. Zero-shot ear cross-dataset transfer for person recognition on mobile devices. **Pattern Recognition Letters**, [S.L.], v. 166, p. 143-150, fev. 2023. Elsevier BV. http://dx.doi.org/10.1016/j.patrec.2023.01.012.

GENG, Lei; ZHANG, Siqi; TONG, Jun; XIAO, Zhitao. Lung segmentation method with dilated convolution based on VGG-16 network. **Computer Assisted Surgery**, [S.L.], v. 24, n. 2, p. 27-33, 12 ago. 2019. Informa UK Limited. http://dx.doi.org/10.1080/24699322.2019.1649071.

GLOROT, Xavier; BENGIO, Yoshua. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. **PMLR** 9:249-256, 2010.

GOODFELLOW, Ian; BENGIO; Yoshua; COURVILLE, Aaron. **Deep learning**. MIT Press, 2016. Disponível em: https://www.deeplearningbook.org/.

GUO, Peng; XUE, Zhiyun; LONG, L. Rodney; ANTANI, Sameer. Cross-Dataset Evaluation of Deep Learning Networks for Uterine Cervix Segmentation. **Diagnostics**, [S.L.], v. 10, n. 1, p. 44, 14 jan. 2020. MDPI AG. http://dx.doi.org/10.3390/diagnostics10010044.

HAMZA, Ameer; KHAN, Muhammad Attique; WANG, Shui-Hua; ALQAHTANI, Abdullah; ALSUBAI, Shtwai; BINBUSAYYIS, Adel; HUSSEIN, Hany S.; MARTINETZ, Thomas Markus; ALSHAZLY, Hammam. COVID-19 classification using chest X-ray images: a framework of CNN-LSTM and improved max value moth flame optimization. **Frontiers In Public Health,** [S.L.], v. 101, n. 1, p. 1-1, 30 ago. 2022. Frontiers Media SA. http://dx.doi.org/10.3389/fpubh.2022.948205.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, [S.L.], v. 143, n. 1, p. 29-36, abr. 1982. Radiological Society of North America (RSNA). http://dx.doi.org/10.1148/radiology.143.1.7063747.

HASSAN, Md Rafiul; ISMAIL, Walaa N.; CHOWDHURY, Ahmad; HOSSAIN, Sharara; HUDA, Shamsul; HASSAN, Mohammad Mehedi. A framework of genetic algorithm-based CNN on multi-access edge computing for automated detection of COVID-19. **The Journal of Supercomputing**, (S.L.), v. 78, n. 7, p. 10250-10274, 21 jan. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11227-021-04222-4.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep Residual Learning for Image Recognition. *In*: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **Anais...** [S.L.], jun. 2016. IEEE. http://dx.doi.org/10.1109/cvpr.2016.90.

JIANG, Xiaoben; ZHU, Yu; CAI, Gan; ZHENG, Bingbing; YANG, Dawei. MXT: a new variant of pyramid vision transformer for multi-label chest X-ray image classification. **Cognitive Computation**, [S.L.], v. 14, n. 4, p. 1362-1377, 3 jun. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12559-022-10032-4.

JUNG, Minhyuk; CHI, Seokho. Human activity classification based on sound recognition and residual convolutional neural network. **Automation in Construction**, [S.L.], v. 114, p. 103177, jun. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.autcon.2020.103177.

KAPOOR, Namrata. Recall, Specificity, Precision, F1-Score and Accuracy. **Numpy Ninja**, Jan 15, 2021. Disponível em: https://www.numpyninja.com/post/recall-specificity-precison-f1-scores-and-accuracy. Acesso em 13 abr. 2023.

KATHAMUTHU, Nirmala Devi; SUBRAMANIAM, Shanthi; LE, Quynh Hoang; MUTHUSAMY, Suresh; PANCHAL, Hitesh; SUNDARARAJAN, Suma Christal Mary; ALRUBAIE, Ali Jawad; ZAHRA, Musaddak Maher Abdul. A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications. **Advances in Engineering Software**, [S.L.], v. 175, p. 103317, jan. 2023. Elsevier BV. http://dx.doi.org/10.1016/j.advengsoft.2022.103317.

KIBRIYA, Hareem; AMIN, Rashid. A residual network-based framework for COVID-19 detection from CXR images. **Neural Computing and Applications**, [S.L.], v. 35, n. 11, p. 8505-8516, 15 dez. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00521-022-08127-y.

KONWER, Aishik; PRASANNA, Prateek. Clinical outcome prediction in COVID-19 using self-supervised vision transformer representations. Medical Imaging 2022: Computer-Aided Diagnosis, **Anais**… [S.L.], 4 abr. 2022. SPIE. http://dx.doi.org/10.1117/12.2612957.

KRISHNAN, Koushik Sivarama; KRISHNAN, Karthik Sivarama. Vision Transformer based COVID-19 Detection using Chest X-rays. *In*: 6Th International Conference on Signal Processing, Computing and Control (ISPCC), **Anais**… [S.L.], 7 out. 2021. IEEE. http://dx.doi.org/10.1109/ispcc53510.2021.9609375.

KRIZHEVSKY, Alex; SUTSKEVER, Sutskever; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, **Anais**... 2012.

LANJEWAR, Madhusudan G.; SHAIKH, Arman Yusuf; PARAB, Jivan. Cloud-based COVID-19 disease prediction system from X-Ray images using convolutional neural network on smartphone. **Multimedia Tools and Applications**, [S.L.], 24 nov. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11042-022-14232-w.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P.. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, [S.L.], v. 86, n. 11, p. 2278-2324, 1998. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/5.726791.

LI, Jingxing; YANG, Zhanglei; YU, Yifan. A Medical AI Diagnosis Platform Based on Vision Transformer for Coronavirus. *In*: 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), **Anais…** [S.L.], 24 set. 2021. IEEE. http://dx.doi.org/10.1109/cei52496.2021.9574576.

LING, C. X.; HUANG, J.; ZHANG, H.. AUC: A statistically consistent and more discriminating measure than accuracy (Conference Paper). *In*: 18th International

Joint Conference on Artificial Intelligence, IJCAI, **Anais...**, 2003; Acapulco; Mexico; 9 August 2003 through 15 August 2003; Code 97871, p. 519-524.

LITJENS, Geert; KOOI, Thijs; BEJNORDI, Babak Ehteshami; SETIO, Arnaud Arindra Adiyoso; CIOMPI, Francesco; GHAFOORIAN, Mohsen; LAAK, Jeroen A.W.M. van Der; VAN GINNEKEN, Bram; SÁNCHEZ, Clara I.. A survey on deep learning in medical image analysis. **Medical Image Analysis**, [S.L.], v. 42, p. 60-88, dez. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.media.2017.07.005.

LIU, Guangyi; LIAO, Yinghong; WANG, Fuyu; ZHANG, Bin; ZHANG, Lu; LIANG, Xiaodan; WAN, Xiang; LI, Shaolin; LI, Zhen; ZHANG, Shuixing. Medical-VLBERT: medical visual language BERT for COVID-19 CT report generation with alternate learning. **IEEE Transactions on Neural Networks and Learning Systems**, [S.L.], v. 32, n. 9, p. 3786-3797, set. 2021. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tnnls.2021.3099165.

LIU, Ze; LIN, Yutong; CAO, Yue; HU, Han; WEI, Yixuan; ZHANG, Zheng; LIN, Stephen; GUO, Baining. Swin Transformer: hierarchical vision transformer using shifted windows. *In*: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), **Anais...** [S.L.], out. 2021. IEEE. http://dx.doi.org/10.1109/iccv48922.2021.00986.

LIU, Jian; SHAO, Haijian; JIANG, Yingtao; DENG, Xing. CNN-Based Hidden-Layer Topological Structure Design and Optimization Methods for Image Classification. **Neural Processing Letters**, [S.L.], v. 54, n. 4, p. 2831-2842, 13 jan. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11063-022-10742-8.

LIU, Mengchen; SHI, Jiaxin; LI, Zhen; LI, Chongxuan; ZHU, Jun; LIU, Shixia. Towards Better Analysis of Deep Convolutional Neural Networks. **IEEE Transactions on Visualization and Computer Graphics**, (S.L.), v. 23, n. 1, p. 91-100, jan. 2017b. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tvcg.2016.2598831.

LIU, Shaobo; SHIH, Frank Y.; ZHONG, Xin. Classification of Chest X-Ray Images Using Novel Adaptive Morphological Neural Networks. **International Journal of Pattern Recognition and Artificial Intelligence**, (S.L.), v. 35, n. 10, p. 2157006, 14 maio 2021. World Scientific Pub Co Pte Lt. http://dx.doi.org/10.1142/s0218001421570068.

LUJÁN-GARCÍA, Juan; YÁÑEZ-MÁRQUEZ, Cornelio; VILLUENDAS-REY, Yenny; CAMACHO-NIETO, Oscar. A Transfer Learning Method for Pneumonia Classification and Visualization. **Applied Sciences**, (S.L.), v. 10, n. 8, p. 2908, 23 abr. 2020. MDPI AG. http://dx.doi.org/10.3390/app10082908.

LUZ, Eduardo; SILVA, Pedro; SILVA, Rodrigo; SILVA, Ludmila; GUIMARÃES, João; MIOZZO, Gustavo; MOREIRA, Gladston; MENOTTI, David. Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images. **Research on Biomedical Engineering**, [S.L.], v. 38, n. 1, p. 149-162, 20 abr. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s42600-021-00151-6.

MA, Yongjun; LV, Wei. Identification of Pneumonia in Chest X-Ray Image Based on Transformer. **International Journal of Antennas and Propagation**, [S.L.], v. 2022, p. 1-8, 1 ago. 2022. Hindawi Limited. http://dx.doi.org/10.1155/2022/5072666.

MAREFAT, Abdolreza; MAREFAT, Mahdieh; JOLOUDARI, Javad Hassannataj; NEMATOLLAHI, Mohammad Ali; LASHGARI, Reza. CCTCOVID: COVID-19 detection from chest X-ray images using compact convolutional transformers. **Frontiers In Public Health**, [S.L.], v. 11, p. 1, 27 fev. 2023. Frontiers Media SA. http://dx.doi.org/10.3389/fpubh.2023.1025746.

MEEDENIYA, Dulani; KUMARASINGHE, Hashara; KOLONNE, Shammi; FERNANDO, Chamodi; DÍEZ, Isabel de La Torre; MARQUES, Gonçalo. Chest X-ray analysis empowered with deep learning: a systematic review. **Applied Soft Computing**, [S.L.], v. 126, p. 109319, set. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.asoc.2022.109319.

MEHBOOB, Fozia; RAUF, Abdul; JIANG, Richard; SAUDAGAR, Abdul Khader Jilani; MALIK, Khalid Mahmood; KHAN, Muhammad Badruddin; HASNAT, Mozaherul Hoque Abdul; ALTAMEEM, Abdullah; ALKHATHAMI, Mohammed. Towards robust diagnosis of COVID-19 using vision self-attention transformer. **Scientific Reports**, [S.L.], v. 12, n. 1, p. 1-1, 26 maio 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-022-13039-x.

MINISTÉRIO DA SAÚDE. **Cadastro Nacional de Estabelecimentos de Saúdes** (CNES). Disponível em: https://cnes.datasus.gov.br/. Acesso em 12 out. 2022.

MONDAL, Arnab Kumar; BHATTACHARJEE, Arnab; SINGLA, Parag; PRATHOSH, A. P.. XViTCOS: explainable vision transformer based COVID-19 screening using radiography. **IEEE Journal of Translational Engineering in Health and Medicine**, [S.L.], v. 10, p. 1-10, 2022. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/jtehm.2021.3134096.

MURPHY, Zachary R.; VENKATESH, Kesavan; SULAM, Jeremias; YI, Paul H.. Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: a comparison of performance, sample efficiency, and hidden stratification. **Radiology: Artificial Intelligence**, [S.L.], v. 4, n. 6, p. 1, 1 nov. 2022. Radiological Society of North America (RSNA). http://dx.doi.org/10.1148/ryai.220012.

NAKAGAWA, Elisa Yumi; FELIZARDO, Katia Romero; FABBRI, Sandra Camargo Pinto Ferraz; FERRARI, Fabiano Cutigi. **Revisão sistemática da literatura em Engenharia de Software**: teoria e prática. 1. ed. São Paulo: GEN LTC, 2017.

NING, Wanshan; LEI, Shijun; YANG, Jingjing; CAO, Yukun; JIANG, Peiran; YANG, Qianqian; ZHANG, Jiao; WANG, Xiaobei; CHEN, Fenghua; GENG, Zhi. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. **Nature Biomedical Engineering**, [S.L.], v. 4, n. 12, p. 1197-1207, 18 nov. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41551-020-00633-5.

NISHIO, Mizuho; KOBAYASHI, Daigo; NISHIOKA, Eiko; MATSUO, Hidetoshi; URASE, Yasuyo; ONOUE, Koji; ISHIKURA, Reiichi; KITAMURA, Yuri; SAKAI, Eiro; TOMITA, Masaru. Deep learning model for the automatic classification of COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy: a multi-center retrospective study. **Scientific Reports**, [S.L.], v. 12, n. 1, p. 1-1, 17 maio 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-022-11990-3.

PAN, Shaoyan; WANG, Tonghe; QIU, Richard L J; AXENTE, Marian; CHANG, Chih-Wei; PENG, Junbo; PATEL, Ashish B; SHELTON, Joseph; A PATEL, Sagar; ROPER, Justin. 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. **Physics In Medicine & Biology**, [S.L.], v. 68, n. 10, p. 105004, 5 maio 2023. IOP Publishing. http://dx.doi.org/10.1088/1361-6560/acca5c.

PARK, Kwangjin; CHOI, Youngjin; LEE, Hongchul. COVID-19 CXR Classification: applying domain extension transfer learning and deep learning. **Applied Sciences**, [S.L.], v. 12, n. 21, p. 10715-1, 22 out. 2022. MDPI AG. http://dx.doi.org/10.3390/app122110715.

PARK, Sangjoon; KIM, Gwanghyun; OH, Yujin; SEO, Joon Beom; LEE, Sang Min; KIM, Jin Hwan; MOON, Sungjun; LIM, Jae-Kwang; YE, Jong Chul. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. **Medical Image Analysis**, [S.L.], v. 75, p. 102299, jan. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.media.2021.102299.

PASCANU, Razvan; MIKOLOV, Tomas; BENGIO, Yoshua. On the difficulty of training Recurrent Neural Networks. **Arxiv**, [S.L.], 2012. ArXiv. http://dx.doi.org/10.48550/ARXIV.1211.5063.

PENG, Lihong; WANG, Chang; TIAN, Geng; LIU, Guangyi; LI, Gan; LU, Yuankang; YANG, Jialiang; CHEN, Min; LI, Zejun. Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet. **Frontiers In Microbiology**, [S.L.], v. 13, p. 1, 23 set. 2022. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2022.995323.

RAFI, Taki Hasan. An ensemble deep transfer-learning approach to identify COVID-19 cases from chest X-ray images. *In*: 2020 IEEE Conference on Computational Intelligence In Bioinformatics And Computational Biology (CIBCB), **Anais...** [S.L.], v. 1, n. 1, p. 1, 27 out. 2020. IEEE. http://dx.doi.org/10.1109/cibcb48159.2020.9277695.

RAHHAL, Mohamad Mahmoud Al; BAZI, Yakoub; JOMAA, Rami M.; ALSHIBLI, Ahmad; ALAJLAN, Naif; MEKHALFI, Mohamed Lamine; MELGANI, Farid. COVID-19 Detection in CT/X-ray Imagery Using Vision Transformers. **Journal of Personalized Medicine**, [S.L.], v. 12, n. 2, p. 310-310, 18 fev. 2022. MDPI AG. http://dx.doi.org/10.3390/jpm12020310.

RAHMAN, Tawsifur; KHANDAKAR, Amith; QIBLAWEY, Yazan; TAHIR, Anas; KIRANYAZ, Serkan; KASHEM, Saad Bin Abul; ISLAM, Mohammad Tariqul; MAADEED, Somaya Al; ZUGHAIER, Susu M.; KHAN, Muhammad Salman. Ex-

ploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. **Computers in Biology and Medicine**, [S.L.], v. 132, p. 104319, maio 2021. Elsevier BV. http://dx.doi.org/10.1016/j.compbiomed.2021.104319.

RAWAT, Waseem; WANG, Zenghui. Deep Convolutional Neural Networks for Image Classification: a comprehensive review. **Neural Computation**, [S.L.], v. 29, n. 9, p. 2352-2449, set. 2017. MIT Press - Journals. http://dx.doi.org/10.1162/neco_a_00990.

SHEFFER, M., *et al*. **O perfil do médico especialista em Radiologia e Diagnóstico por Imagem no Brasil**. São Paulo: CBR, 2019. 179 p. Disponível em: https://cbr.org.br/wp-content/uploads/2022/04/perfil%20medico%20especialista%20rddi%20brasil_portugues_digital_vs%2011-03-20.pdf.

SHEKHAR, Shashank; BANSODE, Adesh; SALIM, Asif. A Comparative study of Hyper-Parameter Optimization Tools. *In*: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, **Anais...**, CSDE 2021. [S.l.], 2021. IEEE. DOI: 10.1109/CSDE53843.2021.9718485.

SHOME, Debaditya; KAR, T.; MOHANTY, Sachi; TIWARI, Prayag; MUHAMMAD, Khan; ALTAMEEM, Abdullah; ZHANG, Yazhou; SAUDAGAR, Abdul. COVID-Transformer: interpretable COVID-19 detection using vision transformer for healthcare. *In:* International Journal of Environmental Research and Public Health, **Anais...,** [S.L.], v. 18, n. 21, p. 11086, 21 out. 2021. MDPI AG. http://dx.doi.org/10.3390/ijerph182111086.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M.. A survey on Image Data Augmentation for Deep Learning. **Journal of Big Data**, [S.L.], v. 6, n. 1, p. 1-1, 6 jul. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40537-019-0197-0.

SILVA, Leandro Nunes de Castro; FERRARI, Daniel Gomes. **Introdução a mineração de dados**. Saraiva Uni, 2017.

SOARES, Eduardo; ANGELOV, Plamen. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. **medRxiv** (2020). Doi: https://doi.org/10.1101/2020.04.24.20078584.

SRIVASTAVA, Gaurav; CHAUHAN, Aninditaa; JANGID, Mahesh; CHAURASIA, Sandeep. CoviXNet: a novel and efficient deep learning model for detection of COVID-19 using chest x-ray images. **Biomedical Signal Processing and Control**, [S.L.], v. 78, p. 103848, set. 2022. Elsevier BV. DOI: 10.1016/j.bspc.2022.103848.

SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **JMLR**, 15(56):1929−1958, 2014. Disponível em: https://jmlr.org/papers/v15/srivastava14a.html.

SUN, Weixuan; QIN, Zhen; DENG, Hui; WANG; Jianyuan; ZHANG, Yi; ZHANG, Kaihao; BARNES, Nick; BIRCHFIELD, Stan; KONG, Lingpeng; ZHONG, Yiran. Vicinity Vision Transformer. **arXiv preprint**, [S.L.], vol. 14, no. 8, August 2015. Disponível em: https://arxiv.org/pdf/2206.10552.pdf.

SZEGEDY, Christian; VANHOUCKE, Vincent; IOFFE, Sergey; SHLENS, Jon; WOJNA, Zbigniew. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **Anais...** [S.L.], jun. 2016. IEEE. http://dx.doi.org/10.1109/cvpr.2016.308.

TAHIR, Anas M.; QIBLAWEY, Yazan; KHANDAKAR, Amith; RAHMAN, Tawsifur; KHURSHID, Uzair; MUSHARAVATI, Farayi; ISLAM, M. T.; KIRANYAZ, Serkan; AL-MAADEED, Somaya; CHOWDHURY, Muhammad E. H.. Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images. **Cognitive Computation**, [S.L.], v. 14, n. 5, p. 1752-1772, 11 jan. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12559-021-09955-1.

THAN, Joel C. M.; THON, Pun Liang; RIJAL, Omar Mohd; KASSIM, Rosminah M.; YUNUS, Ashari; NOOR, Norliza M.; THEN, Patrick. Preliminary Study on Patch Sizes in Vision Transformers (ViT) for COVID-19 and Diseased Lungs Classification. 2021 IEEE National Biomedical Engineering Conference (NBEC), **Anais...** [S.L.], 9 nov. 2021. DOI: http://dx.doi.org/10.1109/nbec53282.2021.9618751.

TIAN, Geng; WANG, Ziwei; WANG, Chang; CHEN, Jianhua; LIU, Guangyi; XU, He; LU, Yuankang; HAN, Zhuoran; ZHAO, Yubo; LI, Zejun. A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt. **Frontiers In Microbiology**, [S.L.], v. 13, p. 1, 4 nov. 2022. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2022.1024104.

TOMPSON, Jonathan; GOROSHIN, Ross; JAIN, Arjun; LECUN, Yann; BREGLER, Christoph. Efficient object localization using Convolutional Networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **Anais...** [S.L.], jun. 2015. IEEE. http://dx.doi.org/10.1109/cvpr.2015.7298664.

TULI, Shikhar; DASGUPTA, Ishita; GRANT, Erin; GRIFFITHS, Thomas L.. Are Convolutional Neural Networks or Transformers more like human vision? **Arxiv**, [S.L.], 2021. http://dx.doi.org/10.48550/ARXIV.2105.07197.

VAN GINNEKEN, Bram. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. **Radiological Physics and Technology**, [S.L.], v. 10, n. 1, p. 23-32, 16 fev. 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12194-017-0394-5.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan; KAISER, Łukasz; POLOSUKHIN, Illia. Attention Is All You Need. In: GUYION, I; *et al.* **Advances in Neural Information Processing Systems**, December:5999–6009, 2017. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85043317328&partnerID=40&md5=3e5a5c2b862c8979ffea845bb707b3c3.

WANG, Tianmu; NIE, Zhenguo; WANG, Ruijing; XU, Qingfeng; HUANG, Hongshi; XU, Handing; XIE, Fugui; LIU, Xin-Jun. PneuNet: deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using vision transformer. **Medical & Biological Engineering & Computing**, [S.L.], v. 61, n. 6, p. 1395-1408, 31 jan. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11517-022-02746-2.

WANG, Sheng; PISCO, Angela Oliveira; MCGEEVER, Aaron; BRBIC, Maria; ZITNIK, Marinka; DARMANIS, Spyros; LESKOVEC, Jure; KARKANIAS, Jim; ALTMAN, Russ B.. Leveraging the Cell Ontology to classify unseen cell types. **Nature Communications**, [S.L.], v. 12, n. 1, 21 set. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41467-021-25725-x.

WEI, Zimian; PAN, Hengyue; LI, Lujun; LU, Menglong; NIU, Xin; DONG, Peijie; LI, Dongsheng. DMFormer: closing the gap between CNN and Vision Transformers. **Arxiv**, [S.L.], 2022. http://dx.doi.org/10.48550/ARXIV.2209.07738.

WOLF, Thomas; DEBUT, Lysandre; SANH, Victor; CHAUMOND, Julien; DELANGUE, Clement; MOI, Anthony; CISTAC, Pierric; RAULT, Tim; LOUF, Remi; FUNTOWICZ, Morgan. Transformers: state-of-the-art natural language processing. *In*: Proceedings of the 2020 Conference on Empirical Methods In Natural Language Processing: System Demonstrations, **Anais...** [S.L.], v. 1, n. 1, p. 1-1, 2020. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.

WORLD HEALTH ORGANIZATION. **WHO Coronavirus (COVID-19) Dashboard**. Disponível em: https://covid19.who.int/. Acesso em 20 maio 2023.

YANG, Xingyi; HE, Xuehai; ZHAO, Jinyu; ZHANG, Yichen; ZHANG, Shanghang; XIE, Pengtao. COVID-CT-Dataset: a CT scan dataset about COVID-19. **Arxiv**, [S.L.], v. 3, n. 1, p. 1-1, 17 jun. 2020. ArXiv. http://dx.doi.org/10.48550/ARXIV.2003.13865.

YAO, Zhenjie; LI, Jiangong; GUAN, Zhaoyu; YE, Yancheng; CHEN, Yixin. Liver disease screening based on densely connected deep neural networks. **Neural Networks**, [S.L.], v. 123, p. 299-304, mar. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.neunet.2019.11.005.

YUAN, Jianjun; WU, Fujun; LI, Yuxi; LI, Jinyi; HUANG, Guojun; HUANG, Quanyong. DPDH-CapNet: a novel lightweight capsule network with non-routing for COVID-19 diagnosis using X-ray images. **Journal of Digital Imaging**, [S.L.], v. 36, n. 3, p. 988-1000, 22 fev. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10278-023-00791-3.

ZANDAMELA, Frank; RATSHIDAHO, Terence; NICOLLS, Fred; STOLTZ, Gene. Cross-dataset performance evaluation of deep learning distracted driver detection algorithms. **Matec Web of Conferences**, [S.L.], v. 370, p. 07002, 2022. EDP Sciences. http://dx.doi.org/10.1051/matecconf/202237007002.

ZHANG, Lei; WEN, Yan. A transformer-based framework for automatic COVID19 diagnosis in chest CTs. *In*: 2021 IEEE/CVF International Conference on Com-

puter Vision Workshops (ICCVW), **Anais…** [S.L.], v. 1, n. 1, p. 1-1, out. 2021. IEEE. http://dx.doi.org/10.1109/iccvw54120.2021.00063.

APÊNDICE 1

Estatísticas Descritivas da AUC

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.853	0.864	0.851	0.859	0.833	0.843	0.847	0.858	0.864	0.846	0.793	0.804	0.846	0.771	0.831
Std. error mean	0.0548	0.0521	0.0563	0.0590	0.0583	0.0590	0.0587	0.0552	0.0560	0.0576	0.0861	0.0541	0.0580	0.0560	0.0549
95% CI mean lower bound	0.727	0.744	0.721	0.723	0.699	0.707	0.711	0.730	0.735	0.713	0.595	0.680	0.712	0.642	0.705
95% CI mean upper bound	0.980	0.985	0.981	0.995	0.968	0.979	0.982	0.985	0.994	0.978	0.992	0.929	0.979	0.900	0.958
Median	0.940	0.920	0.910	0.950	0.910	0.960	0.970	0.910	0.940	0.950	0.960	0.840	0.950	0.730	0.850
Mode	0.990	0.550 [a]	0.910 [a]	0.480 [a]	0.460 [a]	0.980 [a]	0.970	0.750	0.940	0.970	0.990	0.930	0.980	0.500 [a]	0.500 [a]
Standard deviation	0.165	0.156	0.169	0.177	0.175	0.177	0.176	0.166	0.168	0.173	0.258	0.162	0.174	0.168	0.165
Variance	0.0271	0.0244	0.0285	0.0313	0.0306	0.0313	0.0310	0.0274	0.0283	0.0299	0.0666	0.0264	0.0303	0.0283	0.0271
IQR	0.240	0.200	0.180	0.200	0.190	0.250	0.250	0.230	0.230	0.220	0.320	0.230	0.240	0.250	0.250
Range	0.460	0.450	0.490	0.520	0.530	0.490	0.490	0.500	0.500	0.500	0.710	0.480	0.500	0.500	0.500
Minimum	0.530	0.550	0.500	0.480	0.460	0.500	0.510	0.500	0.500	0.500	0.280	0.520	0.500	0.500	0.500
Maximum	0.990	1.00	0.990	1.00	0.990	0.990	1.00	1.00	1.00	1.00	0.990	1.00	1.00	1.00	1.00
Shapiro-Wilk W	0.830	0.844	0.816	0.805	0.848	0.802	0.828	0.832	0.784	0.838	0.804	0.924	0.824	0.946	0.900
Shapiro-Wilk p	0.045	0.064	0.031	0.023	0.071	0.022	0.042	0.047	0.013	0.055	0.023	0.427	0.038	0.646	0.254
25th percentile	0.750	0.780	0.780	0.780	0.770	0.730	0.730	0.750	0.750	0.750	0.670	0.700	0.740	0.660	0.720

Estatísticas Descritivas da AUC

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
50th percentile	0.940	0.920	0.910	0.950	0.910	0.960	0.970	0.910	0.940	0.950	0.960	0.840	0.950	0.730	0.850
75th percentile	0.990	0.980	0.960	0.980	0.960	0.980	0.980	0.980	0.980	0.970	0.990	0.930	0.980	0.910	0.970

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

[a] More than one mode exists, only the first is reported

Estatísticas Descritivas da Acurácia

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.787	0.803	0.781	0.788	0.772	0.847	0.853	0.862	0.870	0.847	0.789	0.772	0.838	0.782	0.841
Std. error mean	0.0535	0.0522	0.0549	0.0600	0.0534	0.0562	0.0547	0.0508	0.0516	0.0546	0.0588	0.0528	0.0511	0.0543	0.0505
95% CI mean lower bound	0.663	0.683	0.655	0.649	0.649	0.717	0.727	0.745	0.751	0.721	0.653	0.650	0.720	0.657	0.725
95% CI mean upper bound	0.910	0.924	0.908	0.926	0.895	0.976	0.979	0.979	0.989	0.973	0.925	0.894	0.956	0.907	0.958
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.740	0.790	0.800	0.850
Mode	0.550 [a]	0.550 [a]	0.510 [a]	0.920	0.540 [a]	0.980 [a]	0.970	0.750	0.940	0.970	0.610 [a]	0.930	0.790	0.540 [a]	0.850
Standard deviation	0.160	0.157	0.165	0.180	0.160	0.169	0.164	0.152	0.155	0.164	0.176	0.158	0.153	0.163	0.151
Variance	0.0257	0.0245	0.0271	0.0324	0.0257	0.0284	0.0269	0.0232	0.0240	0.0268	0.0311	0.0251	0.0235	0.0265	0.0229

Estatísticas Descritivas da Acurácia

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
IQR	0.270	0.280	0.270	0.290	0.290	0.250	0.250	0.230	0.230	0.240	0.330	0.300	0.210	0.270	0.240
Range	0.410	0.440	0.470	0.540	0.430	0.450	0.440	0.450	0.450	0.450	0.430	0.440	0.450	0.460	0.450
Minimum	0.550	0.550	0.510	0.450	0.540	0.540	0.560	0.550	0.550	0.550	0.520	0.560	0.550	0.540	0.550
Maximum	0.960	0.990	0.980	0.990	0.970	0.990	1.00	1.00	1.00	1.00	0.950	1.00	1.00	1.00	1.00
Shapiro-Wilk W	0.869	0.914	0.920	0.890	0.897	0.807	0.831	0.856	0.799	0.838	0.815	0.930	0.890	0.937	0.906
Shapiro-Wilk p	0.121	0.343	0.392	0.199	0.235	0.024	0.045	0.087	0.020	0.054	0.030	0.481	0.201	0.553	0.286
25th percentile	0.650	0.650	0.620	0.630	0.610	0.730	0.730	0.750	0.750	0.730	0.610	0.630	0.770	0.640	0.730
50th percentile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.740	0.790	0.800	0.850
75th percentile	0.920	0.930	0.890	0.920	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

[a] More than one mode exists, only the first is reported

Estatísticas Descritivas da Precisão

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.804	0.816	0.790	0.796	0.784	0.839	0.876	0.834	0.879	0.877	0.798	0.771	0.804	0.780	0.829
Std. error mean	0.0489	0.0506	0.0587	0.0634	0.0528	0.0629	0.0439	0.0774	0.0459	0.0400	0.0594	0.0523	0.0761	0.0554	0.0599
95% CI mean lower bound	0.692	0.699	0.655	0.649	0.663	0.694	0.774	0.656	0.773	0.785	0.661	0.651	0.629	0.652	0.691
95% CI mean upper bound	0.917	0.932	0.925	0.942	0.906	0.984	0.977	1.01	0.985	0.969	0.935	0.892	0.980	0.908	0.967
Median	0.870	0.850	0.840	0.860	0.830	0.960	0.970	0.920	0.940	0.950	0.900	0.710	0.810	0.750	0.860
Mode	0.540 [a]	0.520 [a]	0.410 [a]	0.920	0.470 [a]	0.730 [a]	0.970	0.270 [a]	0.940	0.770 [a]	0.950	0.590 [a]	0.980	0.500 [a]	0.440 [a]
Standard deviation	0.147	0.152	0.176	0.190	0.158	0.189	0.132	0.232	0.138	0.120	0.178	0.157	0.228	0.166	0.180
Variance	0.0216	0.0230	0.0310	0.0362	0.0251	0.0356	0.0174	0.0539	0.0189	0.0144	0.0318	0.0246	0.0521	0.0276	0.0323
IQR	0.220	0.180	0.150	0.180	0.180	0.250	0.250	0.220	0.220	0.200	0.240	0.290	0.220	0.240	0.240
Range	0.420	0.470	0.560	0.620	0.500	0.540	0.300	0.730	0.390	0.290	0.490	0.410	0.730	0.500	0.560
Minimum	0.540	0.520	0.410	0.370	0.470	0.450	0.700	0.270	0.610	0.710	0.460	0.590	0.270	0.500	0.440
Maximum	0.960	0.990	0.970	0.990	0.970	0.990	1.00	1.00	1.00	1.00	0.950	1.00	1.00	1.00	1.00
Shapiro-Wilk W	0.899	0.934	0.876	0.853	0.928	0.791	0.786	0.729	0.813	0.803	0.839	0.881	0.792	0.959	0.867
Shapiro-Wilk p	0.244	0.519	0.144	0.080	0.462	0.016	0.014	0.003	0.029	0.022	0.057	0.161	0.017	0.787	0.113
25th percentile	0.700	0.750	0.740	0.740	0.720	0.730	0.730	0.760	0.760	0.770	0.700	0.640	0.760	0.670	0.730
50th percentile	0.870	0.850	0.840	0.860	0.830	0.960	0.970	0.920	0.940	0.950	0.900	0.710	0.810	0.750	0.860

Estatísticas Descritivas da Precisão

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
75th percenti-le	0.920	0.930	0.890	0.920	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

[a] More than one mode exists, only the first is reported

Estatísticas Descritivas da Sensibilidade

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.787	0.792	0.777	0.786	0.766	0.847	0.847	0.858	0.864	0.846	0.789	0.756	0.810	0.753	0.818
Std. error mean	0.0535	0.0555	0.0575	0.0621	0.0560	0.0562	0.0587	0.0552	0.0560	0.0576	0.0588	0.0573	0.0598	0.0600	0.0568
95% CI mean lower bound	0.663	0.664	0.644	0.642	0.637	0.717	0.711	0.730	0.735	0.713	0.653	0.623	0.672	0.615	0.687
95% CI mean upper bound	0.910	0.920	0.909	0.929	0.895	0.976	0.982	0.985	0.994	0.978	0.925	0.888	0.948	0.892	0.949
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.700	0.780	0.690	0.850
Mode	0.550 [a]	0.650 [a]	0.470 [a]	0.420 [a]	0.490 [a]	0.980 [a]	0.970	0.750	0.940	0.970	0.610 [a]	0.630 [a]	0.980	0.500 [a]	0.500 [a]
Standard deviation	0.160	0.166	0.172	0.186	0.168	0.169	0.176	0.166	0.168	0.173	0.176	0.172	0.179	0.180	0.170
Variance	0.0257	0.0277	0.0297	0.0347	0.0282	0.0284	0.0310	0.0274	0.0283	0.0299	0.0311	0.0295	0.0321	0.0324	0.0290
IQR	0.270	0.280	0.270	0.270	0.290	0.250	0.250	0.230	0.230	0.220	0.330	0.300	0.260	0.280	0.260
Range	0.410	0.490	0.500	0.570	0.470	0.450	0.490	0.500	0.500	0.500	0.430	0.480	0.500	0.500	0.500

Estatísticas Descritivas da Sensibilidade

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Minimum	0.550	0.500	0.470	0.420	0.490	0.540	0.510	0.500	0.500	0.500	0.520	0.520	0.500	0.500	0.500
Maximum	0.960	0.990	0.970	0.990	0.960	0.990	1.00	1.00	1.00	1.00	0.950	1.00	1.00	1.00	1.00
Shapiro-Wilk W	0.869	0.917	0.910	0.891	0.909	0.807	0.828	0.832	0.784	0.838	0.815	0.906	0.898	0.912	0.897
Shapiro-Wilk p	0.121	0.367	0.314	0.204	0.306	0.024	0.042	0.047	0.013	0.055	0.030	0.286	0.242	0.329	0.237
25th percentile	0.650	0.650	0.620	0.640	0.610	0.730	0.730	0.750	0.750	0.750	0.610	0.630	0.720	0.630	0.710
50th percentile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.700	0.780	0.690	0.850
75th percentile	0.920	0.930	0.890	0.910	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

[a] More than one mode exists, only the first is reported

Estatísticas Descritivas do F1-Score

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.764	0.777	0.757	0.787	0.773	0.830	0.832	0.840	0.862	0.846	0.773	0.779	0.800	0.770	0.826
Std. error mean	0.0660	0.0660	0.0668	0.0635	0.0541	0.0686	0.0698	0.0693	0.0578	0.0576	0.0671	0.0658	0.0702	0.0557	0.0552
95% CI mean lower bound	0.612	0.624	0.603	0.640	0.649	0.672	0.671	0.680	0.729	0.713	0.619	0.627	0.638	0.642	0.698

Estatísticas Descritivas do F1-Score

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
95% CI mean upper bound	0.917	0.929	0.911	0.933	0.898	0.988	0.993	1.000	0.996	0.978	0.928	0.931	0.962	0.898	0.953
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.840	0.790	0.720	0.850
Mode	0.410 [a]	0.390 [a]	0.380 [a]	0.380 [a]	0.480 [a]	0.980 [a]	0.970	0.750	0.940	0.970	0.950	0.930	0.980	0.510 [a]	0.500 [a]
Standard deviation	0.198	0.198	0.200	0.190	0.162	0.206	0.209	0.208	0.174	0.173	0.201	0.197	0.211	0.167	0.166
Variance	0.0392	0.0392	0.0401	0.0363	0.0263	0.0424	0.0438	0.0433	0.0301	0.0299	0.0405	0.0390	0.0444	0.0279	0.0274
IQR	0.270	0.290	0.300	0.220	0.260	0.250	0.250	0.230	0.230	0.220	0.330	0.300	0.260	0.250	0.240
Range	0.550	0.600	0.590	0.610	0.490	0.600	0.620	0.650	0.520	0.500	0.520	0.580	0.650	0.490	0.500
Minimum	0.410	0.390	0.380	0.380	0.480	0.390	0.380	0.350	0.480	0.500	0.430	0.420	0.350	0.510	0.500
Maximum	0.960	0.990	0.970	0.990	0.970	0.990	1.00	1.00	1.00	1.00	0.950	1.00	1.00	1.00	1.00
Shapiro-Wilk W	0.880	0.906	0.898	0.876	0.934	0.786	0.800	0.771	0.777	0.838	0.835	0.904	0.861	0.937	0.905
Shapiro-Wilk p	0.155	0.286	0.243	0.144	0.521	0.014	0.021	0.009	0.011	0.055	0.050	0.274	0.099	0.551	0.282
25th per-centile	0.650	0.640	0.590	0.690	0.640	0.730	0.730	0.750	0.750	0.750	0.610	0.630	0.720	0.660	0.730
50th per-centile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.840	0.790	0.720	0.850
75th per-centile	0.920	0.930	0.890	0.910	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

[a] More than one mode exists, only the first is reported