UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ
Doutorado em Ciência da Computação Associação UFMA/UFPI

Luis Henrique Silva Vogado

Classificação Hierárquica de Radiografias do Tórax com Comitê de Redes Neurais Convolucionais

Orientador: Prof. Dr. Rodrigo de Melo Souza Veras
Co-orientador: Prof. Dr. Flávio Henrique Duarte de Araújo

Teresina - PI
Setembro, 2024

Luis Henrique Silva Vogado

Classificação Hierárquica de Radiografias do Tórax com Comitê de Redes Neurais Convolucionais

TESE DE DOUTORADO

Tese apresentada como requisito parcial para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Prof. Dr. Rodrigo de Melo Souza Veras
Co-orientador: Prof. Dr. Flávio Henrique Duarte de Araújo

Teresina - PI
Setembro, 2024

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

V877c Vogado, Luís Henrique Silva.
Classificação hierárquica de radiografias do toráx com comitês de redes neurais convolucionais / Luís Henrique Silva Vogado. -- 2024.
81 f. : color.

Tese (Doutorado) - Universidade Federal do Piauí. Centro de Ciências da Natureza. Programa de Pós-Graduação em Ciência da Computação, Teresina, 2024.
"Orientador: Prof. Dr. Rodrigo de Melo Souza Veras. Coorientador: Prof. Dr. Flávio Henrique Duarte de Araújo"

1. Aprendizado de máquina. 2. Análise de imagens. 3. Diagnóstico computadorizado. 4. Radiografia do toráx. 5. Comitê de classificadores. I. Veras, Rodrigo de Melo Souza. II. Araújo, Flávio Henrique Duarte de. III. Título.

CDD 006.31

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Luis Henrique Silva Vogado

Classificação Hierárquica de Radiografias do Tórax com Comitê de Redes Neurais Convolucionais

A presente Tese de Doutorado foi avaliada e aprovada por banca examinadora composta pelos seguintes membros:

Prof. Dr. Rodrigo de Melo Souza Veras
Orientador
Universidade Federal do Piauí

Prof. Dr. Flávio Henrique Duarte de Araújo
Co-orientador
Universidade Federal do Piauí

Profa. Dra. Andrea Gomes Campos Bianchi
Universidade Federal de Ouro Preto

Prof. Dr. Eduardo James Pereira Souto
Universidade Federal do Amazonas

Prof. Dr. João Dallyson Sousa de Almeida
Universidade Federal do Maranhão

Prof. Dr. Kelson Rômulo Teixeira Aires
Universidade Federal do Piauí

Certificamos que esta é a versão original e final da Tese de Doutorado que foi julgada adequada para obtenção do título de Doutora em Ciência da Computação.

gov.br Documento assinado digitalmente
RODRIGO DE MELO SOUZA VERAS
Data: 24/02/2025 17:39:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo de Melo Souza Veras
Orientador

gov.br Documento assinado digitalmente
ANSELMO CARDOSO DE PAIVA
Data: 24/02/2025 19:59:52-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anselmo Cardoso de Paiva
Coordenador

Teresina - PI, 25 de Setembro de 2024

Aos meus pais Maria das Mercês da Silva e Francisco das Chagas Silva,
por sempre me apoiarem e estarem sempre comigo.

Agradecimentos

Agradeço primeiramente a Deus, a minha mãe Maria das Mercês da Silva que sempre acreditou em mim e me deu todo o apoio que me permite seguir adiante. Também agradeço ao meu pai Francisco das Chagas da Silva (*in memorian*), sem ele minha vida não seria a mesma e todo o apoio e amor que me deu, me trouxeram até este momento. Também pretendo aplicar todos os ensinamentos e orientações aprendidos em vida.

A minha irmã Mariana e minha prima Filomena por estarem sempre comigo e me auxiliarem nas dificuldades do dia a dia. Também agradeço a todos os meus amigos, do ensino médio, da universidade e vida por tornarem essa caminhada mais relaxante e prazerosa.

A meu orientador Rodrigo Veras por todo o suporte e ensinamentos proporcionados desde a primeira iniciação científica. A pesquisa não seria a mesma sem a sua orientação e todos os frutos colhidos em nossa parceria foram os melhores possíveis. Também agradeço todo o suporte do meu coorientador prof. Flávio Henrique Duarte de Araújo e do prof. Pedro de Alcântara dos Santos Neto, assim como os demais professores da UFPI e os meus companheiros do LIMCI e da Maida.health.

Aos que de forma direta ou indiretamente colaboraram na realização deste trabalho, meu muito obrigado!

“Ler é sonhar pela mão de outrem.
Ler mal e por alto é libertarmo-nos da mão que nos conduz.
A superficialidade na erudição é o melhor modo de ler bem e ser profundo.”
(Fernando Pessoa)

Resumo

As radiografias de tórax, ou raios X de tórax, são os exames de imagem mais utilizados diariamente nos hospitais. Responsável por auxiliar na detecção de inúmeras patologias e achados que interferem diretamente na vida do paciente, esse exame é, portanto, fundamental na triagem dos pacientes. O uso de técnicas de visão computacional atrelada ao aprendizado profundo auxiliam na tomada de decisão por parte do médico radiologista, fornecendo uma segunda opinião e consequentemente reduzindo custos operacionais. Diante disso, este trabalho propõe uma metodologia hierárquica e baseada em um comitê de Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) para auxiliar no diagnóstico de exames de raios X de tórax, inicialmente rastreando-os com alta probabilidade de serem normais ou anormais e posterior detecção da patologia. No desenvolvimento da etapa de triagem deste estudo, foi utilizado um conjunto de dados com imagens de raios-X de incidências frontal e lateral. Para a construção do modelo ensemble, foram avaliadas as arquiteturas VGG-16, ResNet50, DenseNet121, MobileNetV2 e NasNetMobile comumente utilizadas na classificação de radiografias de tórax. Um Limiar de Confiança (CTR) foi usado para definir as previsões em Normal de Alta Confiança (HCn), classificação Borderline (BC) ou Anormal de Alta Confiança (HCa). Nos testes realizados, foram alcançados resultados bastante promissores: 54,63 % dos exames foram classificados com alta confiança; dos exames normais, 32% foram classificados como HCn com uma taxa de falsa descoberta (FDR) de 1,68%; e quanto aos exames anormais, 23% foram classificados como HCa com 4,91% taxa de falsas omissões (FOR). Na etapa de detecção de anormalidades, avaliamos as arquiteturas com diferentes pré-treinamentos. No entanto, o melhor resultado foi obtido com a VGG-16 pré-treinada com a base de dados proposta no desenvolvimento da triagem com 352.460 imagens. Foram utilizadas duas bases de dados para avaliar essa etapa, são elas: NIH Chest X-rays 14 com 112.120 imagens, onde foi obtida a AUC média de 0,8484 e CheXpert com 224.316 imagens com AUC média de 0,8736. Os resultados obtidos nas duas etapas desenvolvidas na metodologia proposta são promissores, ressaltando que o modelo pré-treinado na triagem e avaliado na classificação de anormalidades, apresentou o melhor resultado, ressaltando a contribuição da hierarquia na classificação dos exames.

Palavras-chave: Análise de Imagens, Aprendizado de Máquina, Classificação de Imagens, Diagnóstico Assistido por Computador, Radiografia do Tórax, Comitê de Classificadores.

Abstract

Chest radiographs, or chest X-rays, are hospitals' most commonly used imaging tests. This test is essential in patient screening and assists in detecting numerous pathologies and findings that directly interfere with the patient's life. Using computer vision techniques linked to deep learning assists the radiologist in decision-making, providing a second opinion and reducing operating costs. Given this, this work proposes a hierarchical methodology based on a Convolutional Neural Networks (CNNs) committee to assist in diagnosing chest X-ray exams, initially screening them with a high probability of being normal or abnormal and subsequently detecting the pathology. In the development of the screening stage of this study, a dataset with X-ray images of frontal and lateral incidences was used. The VGG-16, ResNet50, DenseNet121, MobileNetV2, and NasNetMobile architectures commonly used in chest radiograph classification were evaluated to build the ensemble model. A Confidence Threshold (CTR) was used to define the predictions in High Confidence Normal (HCn), Borderline (BC), or High Confidence Abnormal (HCa). In the tests performed, very promising results were achieved: 54.63% of the exams were classified with high confidence; of the normal exams, 32% were classified as HCn with a false discovery rate (FDR) of 1.68%; and as for the abnormal exams, 23% were classified as HCa with a false omission rate (FOR) of 4.91%. In the abnormality detection stage, we evaluated the architectures with different pre-training. However, the best result was obtained with VGG-16 pre-trained with the base used in the development of the screening. Two databases were used to evaluate this step: NIH Chest X-rays 14 with 112,120 images, with an average AUC of 0.8484, and CheXpert with 224,316 images with an average AUC of 0.8736. The results obtained in the two steps developed in the proposed methodology are promising, highlighting that the model pre-trained in screening and evaluated in the classification of abnormalities presented the best result, highlighting the contribution of the hierarchy in the classification of exams.

Keywords: Image Analysis, Machine Learning, Image Classification, Computer Aided Diagnosis, Chest X-ray, Ensemble of Classifiers.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Acc	*Acurácia*
ACM	*Comprometimento médio - Average Commitment*
BC	*Borderline*
CAM	*Class Activation Mapping*
CapsNets	*Capsule Networks*
CNN	*Convolutional Neural Network*
CAD	*Computer Aided System*
CTR	*Confidence Threshold*
DICOM	*Digital Imaging and Communications in Medicine*
FDR	*False Discovery Rate*
FN	*False Negative*
FP	*False Positive*
FOR	*False Omission Rate*
FMM	*Fast Marching Method*
HCn	*High Confidence Normal*
HCa	*High Confidence Anormal*
ILSVRC	*ImageNet Large Scale Visual Recognition Challenge*
K	*Índice kappa*
LIME	*Local Interpretable Model-agnostic Explanations*
mDFT	*Deeply Fine-Tuning modificado*
PA	*Posteroanterior*
P	*Precisão*
PACS	*Picture Archiving and Communication System*

PLN	*Processamento de Linguagem Natural*
R	*Recall*
RIR	*Report Interpretation Radiologist*
ROC	*Receiver Operating Characteristic Curve*
RSNA	*Radiological Society of North America*
S	*Especificidade*
TC	*Tomografia Computadorizada*
TN	*True Negative*
TP	*True Positive*

Sumário

1 Introdução

Exames por imagem são ferramentas práticas de apoio ao diagnóstico de doenças. Dentre os inúmeros tipos de exames, a radiografia oferece um baixo custo, além da sua fácil utilização. Devido ao baixo custo, é o exame por imagem mais utilizado no mundo. De acordo com o *National Health Service* da Inglaterra, a radiografia foi realizada aproximadamente 16 milhões de vezes entre abril de 2020 e março de 2021[1]. Além disso, segundo o *National Center for Health Statistics* nos Estados Unidos, as radiografias representam cerca de 85% de todos os exames de imagem realizados em hospitais e consultórios médicos[2]. Presente em regiões de difícil acesso, é comumente utilizado como ferramenta de diagnóstico inicial, permitindo que especialistas observem patologias que são difíceis de rastrear. A radiografia é um dos únicos exames que cobre completamente todas as regiões do corpo humano. Dentre as principais áreas, o tórax apresenta inúmeras patologias que por sua vez estão associadas principalmente aos pulmões e coração. Algumas dessas doenças são: pneumonia, derrame pleural, cardiomegalia e nódulos pulmonares.

A radiografia aplica feixes heterogêneos de raios X sobre o corpo que a partir da sua densidade e composição estrutural determina a quantidade de feixes absorvidos. Na radiografia convencional, os feixes que não são absorvidos são capturados por um detector no aparelho. De acordo com a densidade das partes do corpo, a radiografia irá ilustrar diferentes representações em tons de cinza para os mesmos. Os ossos que possuem a maior densidade são geralmente representados por *pixels* de cor branca, sendo o preto a ausência de densidade.

O desenvolvimento de metodologias computacionais que identifiquem as já mencionadas doenças possibilitaria o desenvolvimento de um sistema de diagnóstico auxiliado por computador (*Computer Aided Systems* - CAD) que pode atuar na detecção e acompanhamento de pacientes. Nos últimos anos, vários trabalhos baseados em modelos de aprendizagem profunda foram aplicados com sucesso em problemas comumente encontrados na área médica (Yanas e Triantaphyllou 2019, Itani, Lecron e Fortemps 2019, Gao et al. 2020). Por exemplo, principalmente devido à pandemia causada pelo COVID-19 (Zhang et al. 2021), pesquisadores usaram técnicas baseadas em modelos de aprendizagem profunda para identificar pacientes infectados com essa doença em tomografia computadorizada (TC) (Xu et al. 2020) e imagens de raios X (Ismael e Şengür 2020, Gomes et al. 2020, Ismael e Şengür 2021).

[1] <https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2020-21-data> (acessado em Setembro de 2021.)

[2] <https://pubmed.ncbi.nlm.nih.gov/24439138/> (acessado em Junho de 2024.)

Os hospitais podem usar sistemas CAD para reduzir custos e ajudar a priorizar cuidados mais extremos, agilizando as linhas de atendimento. No entanto, para fornecer tais benefícios, esses sistemas devem ter uma taxa de erro próxima a 1% do ponto de vista médico (Qin et al. 2018). Além disso, existem outros desafios para a correta previsão de exames, como: obtenção de exames tecnicamente limitados que tendem a dificultar o diagnóstico, problemas relacionados à resolução, posicionamento e existência de exames produzidos em diferentes equipamentos.

Além dos desafios, existem inúmeras patologias ou alterações detectáveis na radiografia de tórax. No entanto, achados anormais específicos localizam-se em regiões próximas ou possuem características semelhantes, dificultando sua correta identificação. Outra adversidade para o desenvolvimento de sistemas CAD é a aquisição de conjuntos de dados devidamente rotulados ou erros do sistema *Picture Archiving and Communication System* (PACS), que é responsável pelo armazenamento e comunicação entre computadores e aparelhos que extraem exames por imagem (Choplin, Boehme e Maynard 1992). Os erros de comunicação podem ocasionar problemas como ruídos nas imagens ou metadados com informações inconsistentes. Diante dos inúmeros desafios para a construção de um sistema CAD que possa generalizar todas as mudanças existentes, propomos uma nova metodologia hierárquica que irá reduzir erros e consequentemente aumentar a precisão do sistema.

1.1 Objetivo Geral

O objetivo geral deste trabalho consiste em propor uma metodologia hierárquica para a classificação de imagens de raios X de tórax dividida em duas etapas, são elas: triagem de exames e identificação de anormalidades. As metodologias empregadas em ambas as etapas são baseadas em técnicas de processamento de imagens e aprendizado profundo para classificar exames de raios X de tórax. Por ser hierárquica, a metodologia proposta consiste em inicialmente realizar a triagem de exames normais e anormais por meio de comitê de classificadores com fatores de confiança, definindo assim os termos: *High Confidence Normal* (HCn) e *High Confidence Anormal* (HCa). Já a segunda etapa segue o fluxo da avaliação de exames, uma vez realizada a triagem e o resultado obtido como anormal, esta etapa irá subclassificar a imagem de acordo com a anormalidade ou achado existente.

1.2 Objetivos Específicos

Para atingir o objetivo geral dessa pesquisa, alguns objetivos específicos devem ser alcançados:

- Avaliar e selecionar as melhores CNNs provenientes da literatura, assim como

definir o conjunto de hiperparâmetros que melhor se adapta ao desempenho necessário para compor o comitê da metodologia proposta e a classificação de anormalidades;

- Validar a metodologia proposta com base nas principais métricas da literatura;
- Encontrar o melhor comprometimento entre os resultados obtidos a partir do erro por classe e da quantidade de imagens classificadas (HCa e HCn);
- Definir o melhor conjunto de fatores de confiança, considerando a obtenção da maior precisão dos resultados;
- Comparar os resultados obtidos pela metodologia com as principais abordagens do estado da arte, avaliando tanto as mesmas métricas, quanto as mesmas bases de dados utilizadas no seu desenvolvimento;

1.3 Principais Contribuições e Produções Científicas

Dentre as principais contribuições deste estudo, podemos destacar:

- Criação de um base de dados heterogênea com diferentes anormalidades e achados que não estão presentes em bancos de dados públicos disponíveis na literatura;
- Nova metodologia de ensemble usando diferentes arquiteturas de CNNs de última geração e diferentes projeções;
- Nova metodologia de avaliação considerando a probabilidade das CNNs gerarem classificações baseadas em fatores de confiança;
- Proposta de uma solução totalmente automática que pode ser facilmente implantada em diferentes unidades médicas, principalmente em hospitais;
- Desenvolvimento de um fluxograma para classificação do exame de raios X que contempla desde a triagem do exame até a identificação das anormalidades, funcionando em diferentes níveis de atendimento ao paciente.

A produção científica do doutorado até então resultou na publicação de artigos científicos em periódicos e congressos. A seguir estão listadas as publicações diretamente relacionadas à pesquisa descrita neste documento:

1. **VOGADO, L. H. S.**; VIEIRA, P. A.; SANTOS NETO, P. A.; LOPES, L. A.; SILVA, G. A.; ARAÚJO, F. H. D.; VERAS, R. M. S.. *Detection of COVID-19 in Chest X-ray*

Images using Transfer Learning with Deep Convolutional Neural Network. In: SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing (SAC), 2021, Virtual. *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (SAC), 2021. v. 36. p. 629-636.

2. **VOGADO, L. H. S.**; ARAÚJO, F. H. D.; SANTOS NETO, P. A.; ALMEIDA, J.; TAVARES, J. M. R. S.; VERAS, R. M. S.. *A Ensemble Methodology for Automatic Classification of Chest X-rays Using Deep Learning*. *Computers in Biology and Medicine*, v. 145, p. 105442, 2022.

1.4 Organização do Trabalho

Os demais capítulos deste trabalho foram organizados em:

- O Capítulo 2 apresenta um resumo dos principais trabalhos encontrados na literatura sobre a classificação de imagens de raios X de tórax.
- O Capítulo 3 descreve as principais ferramentas e procedimentos computacionais utilizados, tais como: CNNs, fator de confiança, comitê, e por fim, as métricas de avaliação dos resultados.
- O Capítulo 4 apresenta as bases de dados utilizadas no desenvolvimento da pesquisa, assim como a metodologia proposta que inclui todo o *pipeline*, desde o pré-processamento das imagens até o resultado final.
- O Capítulo 5 mostra os resultados e apresenta as discussões sobre este estudo.
- Ao final, o Capítulo 6 apresenta as conclusões e sugestões de trabalhos futuros.

2 Trabalhos Relacionados

A literatura atual para o desenvolvimento de metodologias que identifiquem anormalidades/alterações em raios X de tórax é vasta. Segundo Çallı et al. 2021, a maioria dos trabalhos publicados apresentou metodologias baseadas na classificação de imagens. Dentre as principais abordagens estudadas, destacam-se a triagem de exames e sua classificação binária e a predição/detecção de anormalidades específicas. Para tanto, foram levantados na literatura trabalhos alinhados com o objetivo proposto neste estudo.

2.1 Triagem de exames (Normal ou Anormal)

Um dos principais desafios em grandes hospitais é o sistema de triagem de pacientes. As dificuldades geradas pela demora no atendimento aos pacientes podem levar a graves consequências em alguns casos. Assim, diversos autores propuseram metodologias para diferenciar radiografias de tórax saudáveis e anormais. As soluções propostas pretendem agilizar a triagem de pacientes de forma automática. Porém, observamos que na literatura, as metodologias primárias ainda apresentam limitações quanto ao número de alterações encontradas nos exames de tórax (Behzadi-khormouji et al. 2020, Guan e Huang 2020, Chen et al. 2019). Essa dificuldade também está relacionada à disponibilidade de conjuntos de dados na literatura. O conjunto de dados ChestX-ray14 (Wang et al. 2017), com cerca de 112.000 imagens e 13 tipos de patologias, é o mais popular.

Entre os trabalhos analisados, destaca-se o trabalho de Yates, Yates e Harvey 2018, onde foram utilizados dois conjuntos de dados distintos para construir as duas classes que foram classificadas com 94,6% de acurácia. Vale ressaltar que esta abordagem favorece a obtenção de altas taxas de acurácia, pois cada conjunto de dados possui uma característica distinta em termos de resolução, equipamento de imagem e técnicas de pré-processamento aplicadas, o que tende a favorecer o aprendizado da CNN utilizada. O trabalho proposto por Ellis et al. 2020 foi o único dentre os analisados que utilizou ambas as incidências lateral e frontal na classificação dos exames, concatenando ambas as imagens para posterior classificação com uma CNN. Em Dunnmon et al. 2019, os autores avaliaram um conjunto de dados com 216.431 imagens e obtiveram 91% de acurácia utilizando a arquitetura DenseNet121. Em Wong et al. 2020, os autores usaram a concatenação de CNNs para desenvolver uma arquitetura pirâmide de características multimodelo; duas bases de dados foram avaliadas, e foi obtida *Receiver Operating characteristic* (ROC) de 0,821.

Em Tang et al. 2020, os autores propuseram a avaliação de diferentes CNNs para a classificação de exames de raios X frontais em saudáveis ou anormais. As arquiteturas VGG-19, ResNet18, ResNet50, InceptionV3 e DenseNet121 foram avaliadas e não apresentaram diferenças significativas em seus resultados. Além disso, imagens com diferentes resoluções, variando de 256x256, 512x512 e 1024x1024 *pixels*, foram avaliadas e seus resultados, assim como as arquiteturas, não apresentaram variação estatística significativa. Porém, dentre as arquiteturas utilizadas, a que obteve a maior média nas métricas avaliadas foi a ResNet18, com acurácia média de 94,64%.

No trabalho de Dyer et al. 2021, foi apresentado um algoritmo para identificar exames de raios X saudáveis com alto fator de confiança. A abordagem pretendia reduzir a carga de trabalho dos radiologistas, oferecendo resultados com alta probabilidade de serem saudáveis. A metodologia é baseada em um comitê formado pelas arquiteturas DenseNet e EfficientNet B4, que reduziu em até 15% o número de exames avaliados pelo médico com HCn com erro de 2,3%. Os autores usaram 3.887 imagens para desenvolver sua solução.

Em Aktas et al. 2022, a InceptionV3 foi empregada no desenvolvimento de dois desafios da literatura, a detecção de COVID-19 e a triagem de exames normais e anormais. Na triagem de exames, foi utilizada a combinação de quatro bases de dados públicas, de onde foram capturadas 102 mil imagens, sendo 62 mil normais e 40 mil anormais. A acurácia obtida pela metodologia proposta foi de 97% para a detecção de COVID e 84% para a triagem de exames.

Em Dmitry 2023, as redes ResNet-18 e DenseNet-121 foram aplicadas na tarefa de classificação binária de radiografias de tórax, especificamente para diferenciar casos normais de anormais. Utilizando transferência de aprendizado com redes pré-treinadas e um conjunto de dados público de 5.364 imagens, o estudo explorou diferentes métodos de pré-processamento e aumento de dados. A melhor configuração obteve acurácia de 90% com a ResNet-18 e 92% com a DenseNet-121 na detecção de pneumonia.

Ao analisar o problema envolvendo a triagem de exames saudáveis e anormais, observa-se que o melhor desempenho alcançado por diferentes metodologias não foi alcançado pela combinação de uma base de dados heterogênea e com várias anormalidades, o que é crítico para implementação em condições reais de uso, como as encontradas em hospitais comuns e clínicas (Chassagnon et al. 2020).

A Tabela 1 apresenta as metodologias do estado da arte que possuem maior relevância neste trabalho. São elencadas as principais características dos trabalhos de acordo com o ano, arquitetura de CNN empregada na classificação, quantidade total de imagens utilizadas no desenvolvimento da metodologia, disponibilidade de base de dados e o principal resultado apresentado pelos autores.

Tabela 1 – Resumo dos trabalhos relacionados encontrados na literatura para a triagem de exames de raios X normais ou anormais.

Trabalho	Ano	Arquitetura	Qtd. de Imagens	Disponibilidade	Resultados
Yates, Yates e Harvey 2018	2019	InceptionV3 + Fine-tuning	53.149 (N: 1.389, A: 51.760)	Pública	Acc de 94,6%
Dunnmon et al. 2019	2019	DenseNet121	216.431 (N: 45.232, A: 171.199)	Privado	Acc de 91%
Ellis et al. 2020	2020	DenseNet121	6.776 (N: 2.575, A: 4.201)	Privado	Acc de 82%
Wong et al. 2020	2020	VGG16+ResNet50 pyramid	128.886 (N: 64.447, A:64.439)	Pública	AUC de 0,821
Tang et al. 2020	2020	ResNet18	141.617	Pública	Acc de 94,64%
Dyer et al. 2021	2021	Ensemble (DenseNet121 e EffIcientNet b4)	3.887 (N: 2.293, A: 1.594)	Privado	15% de todos os exames com HCn de 97,7%
Aktas et al. 2022	2022	InceptionV3	102.240 (N: 62.108, A: 40.132)	Pública	Acc de 84%
Dmitry 2023	2023	DenseNet121	5.216 (N: 1.341, A: 3.875)	Pública	Acc de 92%

Na Tabela 1, percebe-se que alguns dos trabalhos encontrados na literatura utilizaram conjuntos de dados privados e outros públicos. No entanto, nenhum dos trabalhos utilizaram bases de dados com um bom equilíbrio entre as classes normal e anormal. Isso confirma a demanda por bases de dados que melhor representem ambas as classes. Até mesmo na base Chest-Xray-14, que possui um grande número de imagens com patologias, a classe “sem achados” não garante que os exames sejam normais. Além disso, a heterogeneidade do conjunto de dados utilizado é crucial para a obtenção de um modelo capaz de generalizar diferentes patologias em cenários reais, onde qualquer achado pode aparecer. Essa heterogeneidade está presente apenas nos trabalhos de Tang et al. 2020 e Yates, Yates e Harvey 2018, onde as metodologias propostas foram desenvolvidas e avaliadas usando múltiplas bases de dados. Mesmo assim, os autores não exploraram o uso de imagens de projeção lateral, que podem ser decisivas para um diagnóstico eficiente.

Nos trabalhos de estado da arte encontrados, um fator comum é o uso de técnicas de ajuste fino em CNNs e a avaliação de diferentes arquiteturas. Os autores obtiveram resultados com CNNs consolidadas na literatura, como DenseNet, ResNet e Inception. Isso demonstra que, apesar da proposta de inúmeras CNNs, para implementação em sistemas reais a fim de auxiliar no diagnóstico, as arquiteturas consolidadas são as que tendem a alcançar melhor efetividade.

O ponto crítico é que os autores não apresentaram evidências de que as metodologias propostas podem ser utilizadas em cenários reais. A exceção é a proposta de Dyer et al. 2021, que priorizou a acurácia da metodologia para auxiliar o radiologista

na emissão de laudos normais. Portanto, observando a necessidade de abordagens com respostas confiáveis, arquiteturas consolidadas, bases de dados heterogêneas e uso de todas as incidências, desenvolvemos uma metodologia baseada em comitês com CNNs para auxiliar os radiologistas na produção de predições com alta precisão que preenche todas essas lacunas.

2.2 Detecção de anormalidades

Ao longo dos anos, com o surgimento de novas bases de dados completamente rotuladas de acordo com as principais anormalidades, diversos autores propuseram metodologias voltadas para a detecção dessas alterações. Dentre essas bases de dados, podemos mencionar a ChestX-ray14 (Wang et al. 2017) e a CheXpert (Irvin et al. 2019). Sendo consideradas duas das bases de dados públicas mais utilizadas na detecção de anormalidades e de fácil acesso (Çallı et al. 2021). Elas contam com pelo menos 22 classes quando unidas. O principal objetivo da classificação de anormalidades é orientar o médico radiologista e detectar de forma eficiente as alterações ou patologias, auxiliando na celeridade do atendimento, na precisão do diagnóstico e no tratamento precoce do paciente.

Na Tabela 2 apresentamos metodologias para a classificação de anormalidades considerando as duas bases de dados públicas que serão utilizadas neste trabalho, Chest X-ray-14 e CheXpert. Os trabalhos são classificados de acordo com o dataset, a metodologia de treinamento empregada na arquitetura proposta, técnica de pré-processamento, dimensionalidade da imagem de entrada e resultados.

Tabela 2 – Resumo de trabalhos da literatura que propuseram metodologias de identificação de anormalidades em raios X de tórax.

Trabalho	Treinamento	Pré-processamento	*Input size*	Arquitetura	Resultados (Avg AUC)
Chest x-ray-14					
Wang et al. 2017	*Fine tuning*	Redimensionamento	1024x1024	ResNet50	0,745
Yao et al. 2018	From scratch	Redimensionamento Aumento de dados	512x512	DenseNet	0,761
Guendel et al. 2018	*Fine tuning*	-	1024x1024	DNetLoc (DenseNet-121)	0,807
Mao et al. 2018	*Fine tuning*	Recorte central	224x224x3	DGC-VGGNet16	0,7877
Sirazitdinov et al. 2019	Fine tuning	Aumento de dados	1024x1024	Inception-Resnet-v2	0,808
Baltruschat et al. 2019	*Fine tuning*	-	448x448x1	ResNet50	0,806
DSouza, Abidin e Wismüller 2019	*Fine tuning*	-	340x340	ResNet34	-
Allaouzi e Ahmed 2019	*Feature extraction*	Aumento de dados	224x224	DenseNet-121	0,882
Nugroho 2021	Fine tuning	Redimensionamento	224x224	EfficientNet-B3	0,8313
Karim et al. 2022	*From scratch*	-	1080x1080	*Capsule Networks*	0,867
Kamal et al. 2022	*Fine tuning*	Segmentação semi-supervisionada Redimensionamento Recorte central	512x512	Anatomy-XNet	0,8578
CheXpert					
Allaouzi e Ahmed 2019	*Feature extraction*	Aumento de dados	224x224	DenseNet-121	0,812
Irvin et al. 2019	*Fine tuning*	-	320x320	DenseNet-121	0,8886
Pham et al. 2021	*Fine tuning*	Redução de ruído Normalização	256x256	Comitê (DenseNet-121 + DenseNet-169 + DenseNet-201 + Inception-ResNet-v2 + Xception + NASNetLarge)	0,940
Chong et al. 2022	*Fine tuning*	*Screen Synthesis* *Aumento de dados*	224x224	DenseNet-121	0,906
Kamal et al. 2022	*Fine tuning*	Segmentação semi-supervisionada Redimensionamento Recorte central	512x512	Anatomy-XNet	0,9207

Wang et al. 2017 em 2017, apresentou o dataset Chest x-ray 8, assim como

uma metodologia de rotulação dos laudos da base de dados e uma abordagem de classificação para os exames de raios X. Os autores avaliaram quatro arquiteturas de CNNs, para constituir a abordagem proposta, são elas: AlexNet, GoogLeNet, VGG-16 e ResNet50. Os melhores resultados foram obtidos pela ResNet50, com uma AUC média de 0,745. Além dos resultados obtidos, em Wang et al. 2017, foi apresentado uma divisão oficial do dataset para as classes da base de dados. Essa divisão foi depois adotada por inúmeros autores, como, Yao et al. 2018, Guendel et al. 2018 e Baltruschat et al. 2019.

Em Baltruschat et al. 2019, os autores apresentaram a comparação de diferentes abordagens *multi-label* para a classificação de raios X de tórax. Duas bases de dados públicas foram utilizadas na comparação, são elas: CXR dataset da Open-i e a ChestX-ray14. Para avaliar adequadamente o desempenho da abordagem, os autores realizaram comparações com trabalhos do estado da arte utilizando o k-*fold* com cinco grupos, como apresentado por Wang et al. 2017. Além disso, os autores destacaram que as três principais contribuições do trabalho envolvem: metodologias de treinamento, dimensionalidade da imagem de entrada e a inserção de metadados provenientes dos pacientes e do exame. A arquitetura de CNN aplicada no estudo foi a ResNet50, ela foi refinada e modificada para a inserção dos dados que não são imagens. As imagens foram avaliadas com duas dimensionalidades distintas, sendo elas: 256x256 e 448x448. A melhor configuração foi obtida com entrada de 448x448, apenas um canal de cor e a inclusão das características de sexo, idade, gênero e tipo de aquisição. Os resultados obtidos foram comparados com o estado da arte e ressalta-se o seu desempenho na detecção de enfisema, edema, hernia, consolidação e espessamento pleural.

No trabalho de Karim et al. 2022, a metodologia proposta consiste na utilização de CapsNets (*Capsule Networks*) para a classificação de imagens de raios X de tórax. Foram utilizadas 12 camadas de cápsulas, uma vez que a saída de cada camada foi utilizada como entrada para uma camada convolucional. A base de dados utilizado nos experimentos foi o ChestXray-14 e a comparação com outras arquiteturas do estado da arte apresentou o desempenho superior obtido pelas CapsNets, alcançando uma AUC média de 0,867.

Em Kamal et al. 2022, os autores apresentaram a Anatomy-XNet, uma arquitetura que utiliza informações provenientes de segmentações anatômicas com imagens de raios X para classificar os exames. Um dos desafios encontrados foi a falta de amostras das regiões anatômicas do corpo para realizar a segmentação. Sendo assim, os autores utilizaram uma abordagem semi-supervisionada para gerar máscaras nas bases NIH, CheXpert e MIMIC-CXR. Para compor a metodologia proposta, os autores combinaram um módulo *anatomy-aware attention* e um *backbone* da DenseNet121. Dentre os principais resultados, ressaltamos os valores de AUC de 85,78%, 92,07% e 84,04% nas bases de dados estudadas.

Katona et al. 2024 avaliaram as arquiteturas VGG16, ResNet50, and DenseNet121 para a classificação de radiografias do tórax. Os autores empregaram um algoritmo de otimização em *Hyperband* para definir o melhor conjunto de hiperparâmetros. A metodologia proposta alcançou uma AUC média de 0,8250.

Dos trabalhos analisados no estado da arte, levantamos as seguintes constatações:

- O padrão para treinamento das CNNs mais utilizado foi o ajuste-fino, apenas dois trabalhos utilizaram extração de características, o que demonstra o melhor desempenho das CNNs quando retreinadas;
- Os autores utilizaram uma variação de 1024x1024 até 224x224 para definir a dimensionalidade da imagem de entrada da CNN. No entanto, observamos que os melhores resultados foram alcançados com as dimensionalidades 224x224 e 512x512;
- A DenseNet121 foi, amplamente, a CNN mais empregada dentre os trabalhos do estado da arte;
- Aumento de dados mesmo sendo utilizada como ferramenta para reduzir a probabilidade de ocorrer um overfitting, não foi empregada recorrentemente na literatura de acordo com o estudo. Isso se deve a grande quantidade de imagens disponibilizadas nas bases de dados.

2.3 Considerações Finais

Neste capítulo foram apresentados estudos encontrados na literatura que abordam a classificação de exames de raios X de tórax. A predição das imagens se subdivide de acordo com dois objetivos, o primeiro a triagem de exames, que consiste em determinar se o exame é normal ou anormal e a segunda com a classificação de exames anormais de acordo com seu achado ou patologia. Um resumo dos trabalhos apresentados foi apresentado nas Tabelas 1 e 2. Por meio desta pesquisa é possível verificar uma exclusividade na aplicação de técnicas de aprendizado profundo nos trabalhos mais recentes. Também podemos observar que a maioria dos trabalhos avaliam os seus métodos tanto em bases públicas quanto privadas, uma vez que uma maior heterogeneidade nos dados concede maior robustez e melhor confiabilidade aos resultados.

3 Fundamentação Teórica

Neste capítulo descrevemos as principais técnicas e arquiteturas utilizadas na construção da metodologia proposta. Assim, enfatizamos as abordagens de *deep learning* empregadas e as arquiteturas avaliadas. Ao final, apresentamos a metodologia de avaliação que busca validar os resultados obtidos de acordo com as principais métricas encontradas na literatura.

3.1 Aprendizado por Redes Neurais Convolucionais

As Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) foram apresentadas oficialmente por Yann LeCun (Lecun et al. 1998) e Fukushima (Fukushima 1988). No entanto, se tornaram populares por meio do *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC). Aplicadas na resolução de diferentes desafios envolvendo classificação de imagens, reconhecimento e detecção de objetos, substituindo assim as técnicas mais tradicionais que envolvem a extração de características com descritores e classificação de vetores de atributos.

As CNNs são geralmente constituídas por três tipos principais de camadas, são elas: convolucionais, *pooling*, totalmente conectadas. As camadas convolucionais possuem filtros que são convoluídos com a imagem de entrada e extraem uma representação das suas principais características. Traçando um paralelo as redes neurais mais simples, podemos comparar os filtros convolucionais aos pesos que serão reajustados pelo algoritmo de retropropagação do erro. Na Figura 1, ilustramos um exemplo da operação de convolução realizada pela CNN. Para CNNs mais profundas, os filtros das camadas iniciais tendem a extrair características generalistas como borda, textura e cor (Vogado et al. 2021). Já as camadas consideradas mais profundas tendem a extrair características específicas da imagem e auxiliam na melhor separação entre as classes.

As camadas de *pooling* sucedem as camadas convolucionais pois seu principal objetivo é reduzir a complexidade dos mapas de características gerados a partir das convoluções. Para exemplificar, uma camada convolucional com dimensionalidade 12 x 12 x 3 (432 *pixels*) e será convoluída com uma camada que possui 64 filtros com dimensionalidade 5 x 5, terá como saída mapas de características com 8 x 8 x 64 (4096 *pixels*), ou seja, a complexidade aumentará em quase dez vezes. Dentre as camadas de *pooling*, o *maxpooling* irá agrupar e irá selecionar apenas o *pixel* de maior valor para representar aquele conjunto de *pixels*.

Assim como nas redes neurais mais simples, as camadas densas ou totalmente

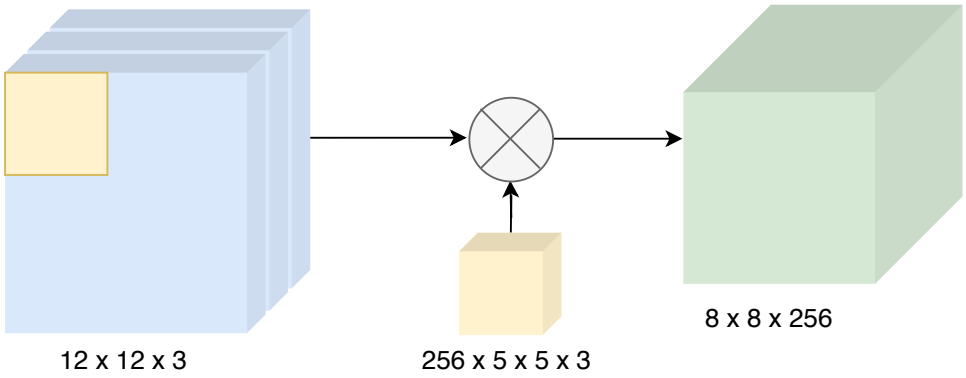


Figura 1 – Operação de convolução realizada nas camadas convolucionais de uma CNN.

conectadas são utilizadas ao final da CNN e reúnem as características de alto nível em uma nova representação, dessa vez por meio de vetores de atributos que são utilizados pela camada de classificação. A operação de *flattening* é aplicada para transformar os mapas de características das camadas convolucionais em um vetor unidimensional que funciona como entrada para as camadas densas. A última camada de classificação é também uma camada densa, no entanto, possui uma função de ativação que irá modelar operações lineares em não-lineares, permitindo que os dados sejam classificados de acordo com o número de classes. Para problemas multiclasses, a *softmax* é a função de ativação preterida, no entanto, para problemas binários e multilabel, a *sigmoid* é a mais utilizada (Goodfellow, Bengio e Courville 2017).

3.1.1 Desafios no uso de CNNs

As CNNs são arquiteturas com grande capacidade de aprender e generalizar múltiplos padrões. No entanto, essas características produzem também alguns desafios que devem ser atendidos para se atingir esses ganhos. O primeiro é a necessidade de grande quantidade de dados para realizar o treinamento das mesmas. Por conta da sua profundidade e a presença de inúmeros parâmetros provenientes das camadas convolucionais e totalmente conectadas, a quantidade necessária de imagens pode facilmente passar de 100 mil para determinados problemas. Nos casos em que a quantidade é insuficiente, geralmente ocorre o *overfitting* ou sobreajuste nos dados, em que as CNNs irão decorar o conjunto de treinamento, mas não conseguirão representar o aprendizado no conjunto de teste (Yosinski et al. 2014).

Além da quantidade de dados, outro desafio é o custo computacional atrelado a quantidade de operações realizadas pelas CNNs. Por conta do grande número de parâmetros e dados, é necessário o uso de tecnologias como *Graphics Processing Units* (GPUs) para a execução das operações em tempo hábil.

Mesmo entendendo que esses dois desafios estão bem estabelecidos na literatura, existem metodologias que utilizam-se dos recursos das CNNs para contorná-los. Algumas

dessas metodologias são a transferência de aprendizado e o ajuste-fino.

3.1.2 Transferência de aprendizado e ajuste fino

As técnicas baseadas em *deep learning* vem sendo aplicadas ao longo dos anos em soluções para os mais diversos problemas. Na literatura, as CNNs têm sido aplicadas para desenvolver soluções que envolvem o diagnóstico de imagens médicas. Essas redes possuem alta generalização, sobrepondo técnicas tradicionais comumente apresentadas na literatura (Çallı et al. 2021). Dentre as principais técnicas envolvendo CNNs, podemos destacar a técnica de transferência de aprendizado, onde uma CNN é pré-treinada em um conjunto de dados genérico é então utilizada como base para mudanças na arquitetura por meio de ajuste fino para um novo problema. Na Figura 2, apresentamos o exemplo da técnica de transferência de aprendizado por ajuste fino com a base de dados ImageNet.

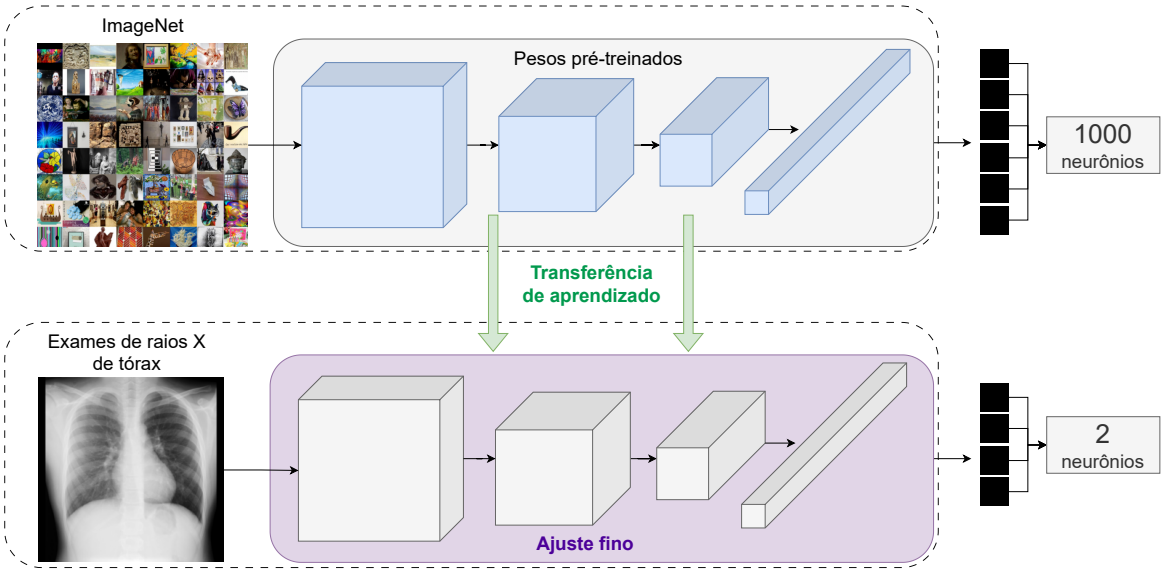


Figura 2 – Fluxograma ilustrando a transferência de aprendizado e ajuste fino utilizados no desenvolvimento do modelo proposto.

Neste trabalho, utilizamos a técnica de ajuste fino para treinar a arquitetura proposta. Em Vogado et al. (Vogado et al. 2021), os autores utilizaram diferentes técnicas de ajuste fino para desenvolver uma arquitetura que classifique corretamente lâminas de sangue com ou sem leucemia. Dentre as abordagens apresentadas, o *modified Deeply Fine-Tuning* (mDFT) consiste em ajustar toda a arquitetura CNN e reajustar as camadas totalmente conectadas. Esta técnica de ajuste fino alcançou os melhores resultados para o problema dado considerando desafios como o tamanho do conjunto de dados e o problema apresentado. Portanto, utilizamos o mDFT para treinar as arquiteturas utilizadas na metodologia proposta.

3.1.3 CNNs avaliadas

A definição da técnica de ajuste fino a ser utilizada foi apenas um passo no desenvolvimento da metodologia proposta. Outro passo fundamental foi a definição da CNN base para o desenvolvimento da arquitetura proposta.

Ao longo dos anos, várias arquiteturas de CNNs foram propostas para resolver problemas de visão computacional e aprendizado de máquina. Essas arquiteturas podem generalizar diferentes conjuntos de dados e fornecer resultados precisos. A partir do ILSVRC (Russakovsky et al. 2015), foram apresentadas arquiteturas com características distintas como profundidade, número de parâmetros e camadas convolucionais. Entre eles, podemos destacar VGG-16 (Simonyan e Zisserman 2014), ResNet50 (He et al. 2016), DenseNet121 (Huang et al. 2017), MobileNetV2 (Howard et al. 2017) e NasNetMobile (Zoph et al. 2018). Em trabalhos que apresentam metodologias para classificação de imagens de radiografia de tórax, esses três modelos de deep learning são os mais utilizados (Çallı et al. 2021). Na Tabela 3, são indicadas as principais características dessas três arquiteturas.

Tabela 3 – Características dos modelos de *deep learning*.

Arquitetura	Profundidade topológica	Qtd. de parâmetros	Ano	Acurácia Top-1	Acurácia Top-5
VGG-16	23	138M	2014	71,3%	90,1%
ResNet50	168	25M	2015	74,9%	92,1%
DenseNet121	159	8.1M	2017	75%	92,3%
MobileNetV2	105	3.5M	2017	71.3%	90.1%
NasNetMobile	389	5.3M	2018	74,4%	91,9%

3.1.3.1 VGG-19

Proposta por Simonyan e Zisserman 2014, esta arquitetura apresenta duas versões principais: os modelos VGG-16 e VGG-19. A diferença fundamental entre eles reside na quantidade de camadas convolucionais. Uma característica notável dessas CNNs é a utilização de múltiplos filtros de menor dimensionalidade, que substituem um único filtro de maior dimensionalidade. Isso resulta na redução do número de parâmetros, levando a uma diminuição do custo computacional associado.

Os autores realizaram diversas configurações através de um estudo de ablação invertido[1] com o objetivo de otimizar a estrutura da VGGNet (**?**). Inicialmente, foi criada uma arquitetura com apenas 11 camadas convolucionais, seguida pela adição da operação de *local response normalization* (LRN), resultando na VGG-11 LRN. Posteriormente, foram conduzidos experimentos com 13 camadas convolucionais, culminando na VGG-13, e finalmente, desenvolveram-se a VGG-16 e a VGG-19, que

[1] O termo “estudo de ablação” refere-se ao procedimento em que partes específicas das redes neurais são removidas, visando uma melhor compreensão do comportamento geral da rede.

aumentaram o número de camadas convolucionais. Contudo, a VGG-19 não conseguiu superar a taxa de erro da VGG-16. Assim, os autores decidiram encerrar o estudo de ablação.

A VGG-16 compreende 13 camadas convolucionais organizadas em cinco blocos, além de três camadas totalmente conectadas, totalizando 16 camadas treináveis. Os dois primeiros blocos contêm apenas duas camadas convolucionais, enquanto os demais apresentam três. Existe uma camada de *maxpooling* entre cada bloco convolucional. Por fim, há duas camadas totalmente conectadas com 4096 unidades cada e uma camada de saída com a função de ativação *softmax*. No total, essa arquitetura possui aproximadamente 138 milhões de parâmetros.

Comparada à VGG-16, a VGG-19 conta com quatro camadas por bloco convolucional (em vez de três), totalizando cerca de 143 milhões de parâmetros. A Figura 3 ilustra a topologia geral de ambas as arquiteturas analisadas neste estudo.

3.1.3.2 ResNet50

Proposta por He et al. 2016, a Rede Neural Residual foi criada para enfrentar o problema do desaparecimento do gradiente, ou *vanishing gradient*, que ocorre quando se adicionam muitas camadas em um modelo sequencial. Com o aumento do número de camadas, o gradiente decresce à medida que é propagado de volta na topologia, devido ao excesso de operações. Como consequência, o aprendizado torna-se mais lento, levando à saturação no desempenho da CNN e, eventualmente, à sua degradação.

Para contornar esse problema, a arquitetura ResNet utiliza blocos residuais que permitem o *skip connection* entre a entrada e a saída de cada bloco. Como a saída representa apenas uma modificação da entrada, os mapas residuais são mais fáceis de otimizar, o que previne a degradação resultante do grande número de camadas. A ResNet conquistou a competição ILSVRC-2015, apresentando uma taxa de erro de apenas 3,6%, inferior à taxa de erro dos seres humanos, que varia entre 5% e 10%. He et al. 2016 propuseram cinco variações da ResNet, com diferentes profundidades: 18, 34, 50, 101 e 152 camadas treináveis. Essas arquiteturas alcançaram taxas de erro menores em comparação com os modelos mais populares na época, como a VGG-16 e a GoogLeNet.

Neste trabalho, avaliamos o uso da ResNet50, que possui uma estrutura com 168 camadas, sendo 50 delas treináveis. A Figura4 apresenta a versão da ResNet com 34 camadas treináveis.

3.1.3.3 DenseNet121

Proposta por Huang et al. 2017, a DenseNet (ou Rede Neural Densa) foi desenvolvida para aprimorar a eficiência do fluxo de gradientes e evitar o problema do desaparecimento do gradiente, semelhante à ResNet, mas com uma abordagem

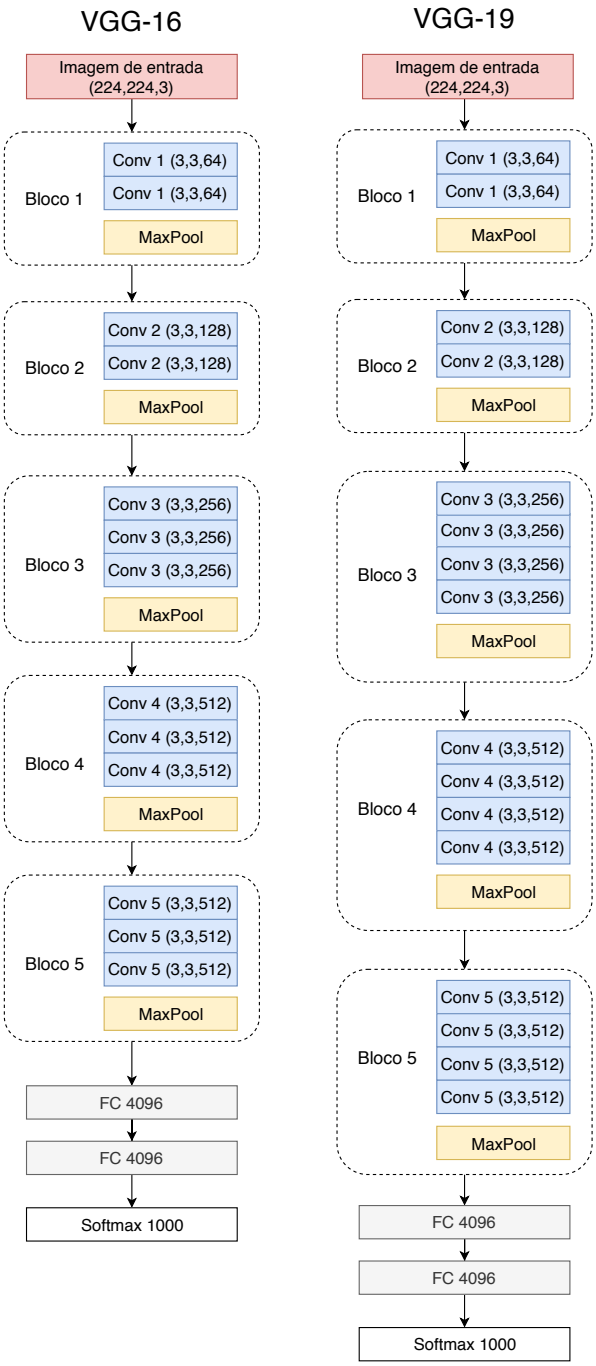


Figura 3 – Estrutura das arquiteturas VGG-16 e VGG-19.

diferenciada. Ao invés de pular conexões, a DenseNet conecta diretamente cada camada a todas as camadas seguintes dentro de um bloco denso. Isso resulta em uma propagação mais eficiente dos gradientes e permite a reutilização dos mapas de características, reduzindo a redundância de informações.

A arquitetura DenseNet é composta de blocos densos, onde cada camada recebe como entrada os mapas de características de todas as camadas anteriores e passa seus próprios mapas para todas as camadas subsequentes. Esse esquema gera uma rede mais estreita e leve, exigindo menos parâmetros e oferecendo vantagens em termos de

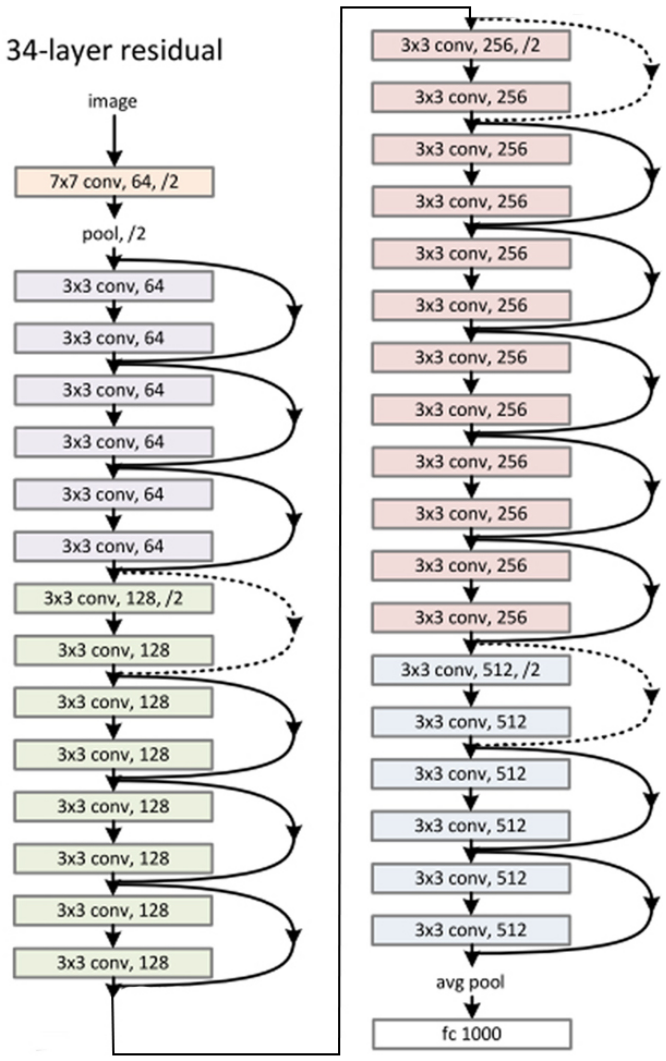


Figura 4 – Residual Network com 34 camadas treináveis. Fonte: (He et al. 2016)

eficiência computacional e memória, especialmente em arquiteturas mais profundas. Como resultado, o modelo utiliza menos parâmetros que redes convencionais como a ResNet, enquanto mantém, ou até melhora, o desempenho.

A DenseNet alcançou resultados notáveis em benchmarks de reconhecimento visual, competindo de forma vantajosa contra arquiteturas como a ResNet e a Inception. Dentre as variantes da DenseNet, a DenseNet121, utilizada neste trabalho, possui uma configuração com 121 camadas e 8M de parâmetros, o que é relativamente leve, considerando o número de conexões e a profundidade.

Observamos que o número de parâmetros necessários para treinar as arquiteturas foi reduzido ao longo dos anos, de mais de 138 milhões para 8 milhões. No entanto, a profundidade topológica aumentou de 23 para 168 e depois para 159 camadas. Nos experimentos, aplicamos mDFT com essas três arquiteturas em projeções de imagem frontal e lateral.

3.1.3.4 MobileNet

A MobileNet, proposta por Howard et al. 2017, foi desenvolvida para otimizar o desempenho de redes neurais profundas em dispositivos com restrições de processamento e memória, como smartphones e dispositivos IoT. A principal inovação da MobileNet é o uso de convoluções separáveis em profundidade (ou *depthwise separable convolutions*), uma técnica que divide a operação de convolução padrão em duas etapas: convoluções em profundidade, aplicadas separadamente a cada canal de entrada, e convoluções ponto a ponto (*pointwise convolutions*), que combinam as saídas desses canais. Esse processo reduz significativamente o número de operações e parâmetros necessários, tornando a rede mais leve e eficiente.

Combinada com hiperparâmetros ajustáveis de largura (*width multiplier*) e resolução (*resolution multiplier*), a MobileNet permite controlar a precisão e a eficiência do modelo, adaptando-o às capacidades de diferentes dispositivos. Em testes de *benchmark*, a MobileNet mostrou-se competitiva com redes maiores em termos de precisão, enquanto é muito mais eficiente em relação ao uso de memória e consumo de energia. Essas características tornaram a MobileNet amplamente utilizada em aplicações de visão computacional móveis e em tempo real, onde eficiência é essencial. Na Figura 5 apresentamos a arquitetura da MobileNet.

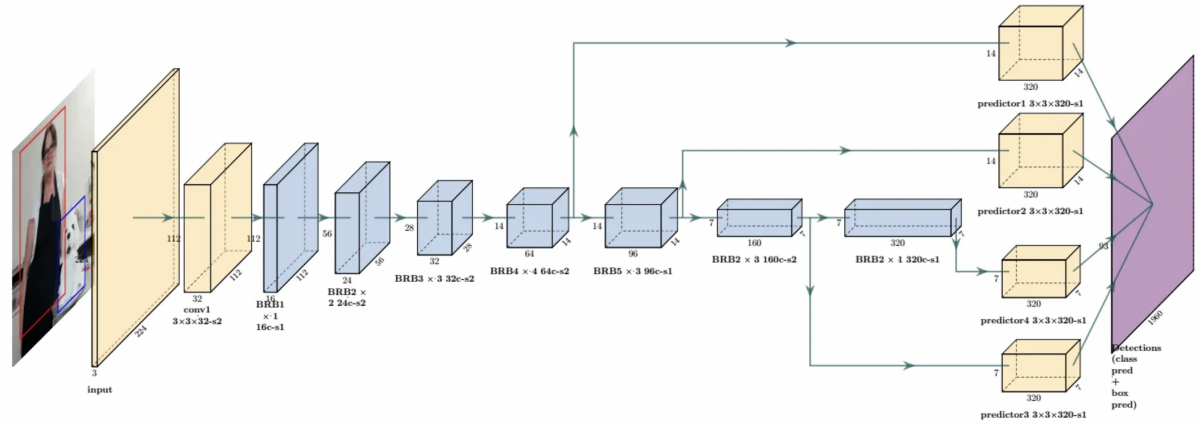


Figura 5 – Arquitetura da MobileNet. Fonte: (Howard et al. 2017)

3.1.3.5 NasNetMobile

A NasNetMobile, introduzida por Zoph et al. 2018, é uma arquitetura de rede neural projetada especificamente para dispositivos móveis e sistemas com restrições de recursos. Desenvolvida com base na busca neural automatizada de arquiteturas (NAS, do inglês *Neural Architecture Search*), a NasNetMobile usa algoritmos de aprendizado automático para descobrir e otimizar sua estrutura, balanceando desempenho e eficiência de forma mais eficaz do que redes convencionais projetadas manualmente.

A NasNetMobile é composta por módulos básicos otimizados para encontrar combinações ideais de blocos de convolução, maximizando a precisão enquanto reduz a complexidade computacional. Esses blocos modulares incluem camadas que conectam diretamente suas saídas às camadas posteriores, o que auxilia na manutenção do fluxo de gradientes em arquiteturas mais profundas. Essa estrutura flexível permite que o modelo alcance uma alta taxa de acurácia em *benchmarks* de classificação de imagens, com um número relativamente pequeno de parâmetros em comparação com outras arquiteturas móveis, como a MobileNet.

Devido ao seu design otimizado, a NasNetMobile é frequentemente usada em dispositivos móveis para aplicações de visão computacional que exigem alta eficiência e precisão, sem comprometer significativamente o desempenho em dispositivos com baixa capacidade de processamento. Na Figura 6 apresentamos a arquitetura da NasNetMobile.

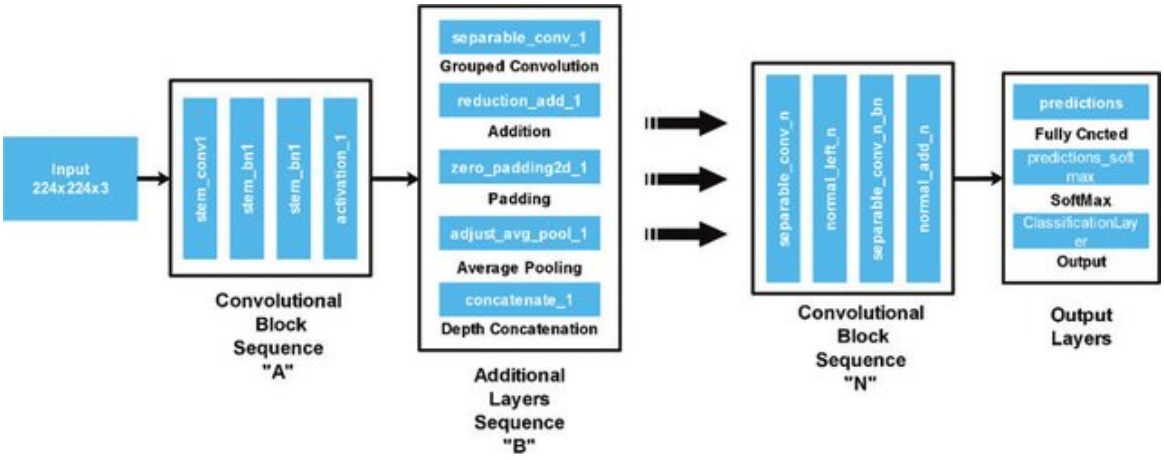


Figura 6 – Arquitetura da NasNetMobile. Fonte: (Zoph et al. 2018)

3.2 Métricas de avaliação

As métricas geralmente utilizadas para avaliar as metodologias de auxílio diagnóstico são baseadas na matriz de confusão. Com base nessa matriz, é possível visualizar e avaliar o desempenho de um algoritmo de previsão por meio da verificação das predições. Assim, para problemas binários, podemos representar a matriz de confusão de acordo com os seguintes valores: verdadeiro positivo (*True Positive* - TP), falso positivo (*False Positive* - FP), falso negativo (*False Negative* - FN) e verdadeiro negativo (*True Negative* -TN). Na Tabela 4, demonstramos como a matriz de confusão é construída a partir das classes positivas (1) e negativas (0).

No problema abordado neste estudo, a classe normal é representada como negativa e a anormal como positiva. Assim, TP representa o que foi classificado corretamente como anormal e FP o que é normal, mas classificado como anormal, TN representa as imagens classificadas corretamente como anormais e FN as

Tabela 4 – Funcionamento da matriz de confusão.

	Atualmente positivo (1)	**Atualmente negativo (0)**
Predito positivo (1)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Predito negativo (0)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

imagens anormais classificadas como normais. A partir desses valores, podemos calcular as métricas de avaliação. Avaliamos as seguintes métricas para selecionar os melhores modelos dos experimentos realizados: acurácia (Acc), precisão (P), recall (R), especificidade (S), índice kappa (K) e área sob a curva ROC (AUC).

Além das métricas apresentadas no estado da arte, também demonstramos métricas alinhadas com os resultados das classificações do comitê. A primeira é a taxa de falsa descoberta (*False Discovery Rate* - FDR) (3.1) e a segunda é a taxa de falsa omissão (*False Omission Rate* - FOR) (3.2):

$$FDR = \frac{FP}{TP + FP}, \tag{3.1}$$

$$FOR = \frac{FN}{TN + FN}. \tag{3.2}$$

Essas métricas são calculadas de acordo com a matriz de confusão resultante do comitê, concordando com o número de respostas HCn e HCa.

Além das métricas FDR e FOR, propomos duas métricas para selecionar os melhores comitês. Como o objetivo principal é aumentar a porcentagem de respostas com alta confiança (HC) e reduzir os erros FDR e FOR, propomos a métrica denominada Comprometimento ou *Commitment* (CM) que representa a média ponderada do número de respostas e o erro obtido naquela classe, que foram definidos de acordo com as Equações 3.3 e 3.4.

$$CM_n = (0.4 * HC_n) + (0.6 * FDR), \tag{3.3}$$

$$CM_a = (0.4 * HC_a) + (0.6 * FOR), \tag{3.4}$$

onde a representa a classe anormal e n a classe normal. Definimos 0,4 como peso para HC_n e HC_a e 0,6 para os respectivos erros por classe: FDR e FOR. Desta forma, o erro tem uma influência substancial na decisão da melhor abordagem de conjunto. A partir da seleção do melhor comprometimento para cada classe e com as abordagens definidas,

ainda é necessário tomar a decisão final entre as combinações de diferentes conjuntos. Para isso, utilizou-se a média ponderada dos valores de Comprometimento (ACM) para cada classe, dando maior peso à classe normal, com 0,6 e 0,4 para a classe anormal.

3.3 Considerações Finais

Neste capítulo apresentamos a fundamentação teórica aplicada a metodologia proposta. Para classificar as imagens, recorremos a utilização de CNNs, que extraem e aprendem características por meio de camadas convolucionais, *max pooling* e densas para prover a melhor generalização do conjunto de dados. Além disso, introduzimos o conceito de transferência de aprendizado, hoje muito empregado pelo estado da arte no desenvolvimento de soluções com CNNs e o ajuste-fino. Por fim, ilustramos as principais métricas que serão utilizadas para avaliar a metodologia proposta e comparar seus resultados com a literatura.

4 Materiais e Métodos

Neste capítulo são apresentadas as principais bases de dados utilizadas no desenvolvimento da metodologia proposta, enfatizando tanto a etapa de triagem quanto a etapa de classificação de anormalidades. Além disso, detalhamos a metodologia proposta e seus principais componentes, como a etapa de pré-processamento, comitê de classificadores e a classificação multilabel.

4.1 Bases de dados

A necessidade de grandes quantidades de dados para treinar metodologias baseadas em *deep learning* é bem conhecida. No entanto, existem outros desafios para o desenvolvimento de ferramentas de auxílio diagnóstico para o uso no mundo real. Um dos principais problemas são a rotulação e o pré-processamento dos dados. Entre os conjuntos de dados públicos encontrados e utilizados nos trabalhos do estado da arte, existem imprecisões quanto à rotulação de dados, principalmente quando são consideradas várias classes. Outro fato que costuma afetar bases de dados públicas é a sobreposição e dependência entre as classes estudadas (Çallı et al. 2021).

Para o desenvolvimento da metodologia hierárquica proposta, dividimos as bases de dados utilizadas de acordo com as etapas necessárias para a classificação dos exames. Sendo assim, para a triagem de exames e detecção de anormalidades combinadas, foram avaliadas cinco bases de dados.

4.1.1 Base de dados proposta

Um desafio encontrado no estado da arte quanto a triagem de exames normais e anormais é a necessidade de bases de dados focadas neste objetivo. Observamos que diversos autores utilizaram a combinação de pelo menos duas bases, sendo uma representando a classe anormal e outra a classe normal para atender as necessidades do problema a ser resolvido. Tendo em vista esse desafio, neste trabalho, apresentamos uma nova base de dados constituída de imagens rotuladas como normais e anormais e coletada em 84 hospitais brasileiros, totalizando 217.302 exames.

O conjunto de dados possui ao todo 352.460 imagens anônimas e as resoluções das imagens variam de 727x692 a 4892x4020 *pixels*. Os exames foram obtidos no formato DICOM e convertidos para posterior processamento. O atributo existente no DICOM, Interpretação Fotométrica veio com o parâmetro *Monochrome*1 para alguns exames, fazendo com que o menor valor de pixel seja exibido pela cor branca, contrastando a maioria dos exames que vieram com *Monochrome*2. Portanto, as imagens

foram ajustadas para satisfazer o padrão *Monochrome*2 por meio da obtenção da negativa da imagem (sua inversão nos valores dos *pixels*).

Os exames foram obtidos de acordo com a incidência frontal (anteroposterior/ posteroanterior) e lateral. Vale ressaltar que nem sempre os protocolos exigem que o médico solicite a incidência lateral. Sendo assim, todos os exames possuem pelo menos uma imagem frontal e uma ou nenhuma imagem lateral, em um total de 224.042 imagens de incidência frontal e 128.418 de incidência lateral. Na Tabela 5, são apresentadas as principais características da base de dados utilizada. A Figura 7 apresenta exemplos de raios X de tórax com sua incidência frontal e lateral.

Utilizamos duas metodologias de rotulação para os exames incluídos na base de dados. A primeira foi o *Report Interpretation Radiologist* (RIR) (Çallı et al. 2021), onde o especialista analisa o laudo médico e o classifica de acordo com o conteúdo. Essa metodologia foi utilizada para todos os exames da classe normal e para parte dos exames anormais. A segunda metodologia de rotulação utilizada foi o *Report Parsing* (RP) (Çallı et al. 2021), onde aplicamos técnicas automáticas para classificar os relatórios. Neste trabalho, utilizamos um dicionário coletado com auxílio dos radiologistas com termos que denotam anormalidades ou achados. Além disso, a busca foi realizada em laudos que não foram classificados como normais na primeira metodologia. Um fato a ser destacado é a definição do termo “anormal”. Neste caso, como pretendemos simular o conhecimento médico mais preciso, usamos qualquer termo que fugisse da normalidade como anormal.

Tabela 5 – Detalhes sobre a base de dados proposta utilizada nos experimentos.

	Frontal	**Lateral**	**Rótulos**	**Método de rotulação**	**Classificação binária**
Base de dados proposta	224.042	128.418	2	RP, RIR	Normal: 236.350 Anormal: 116.110

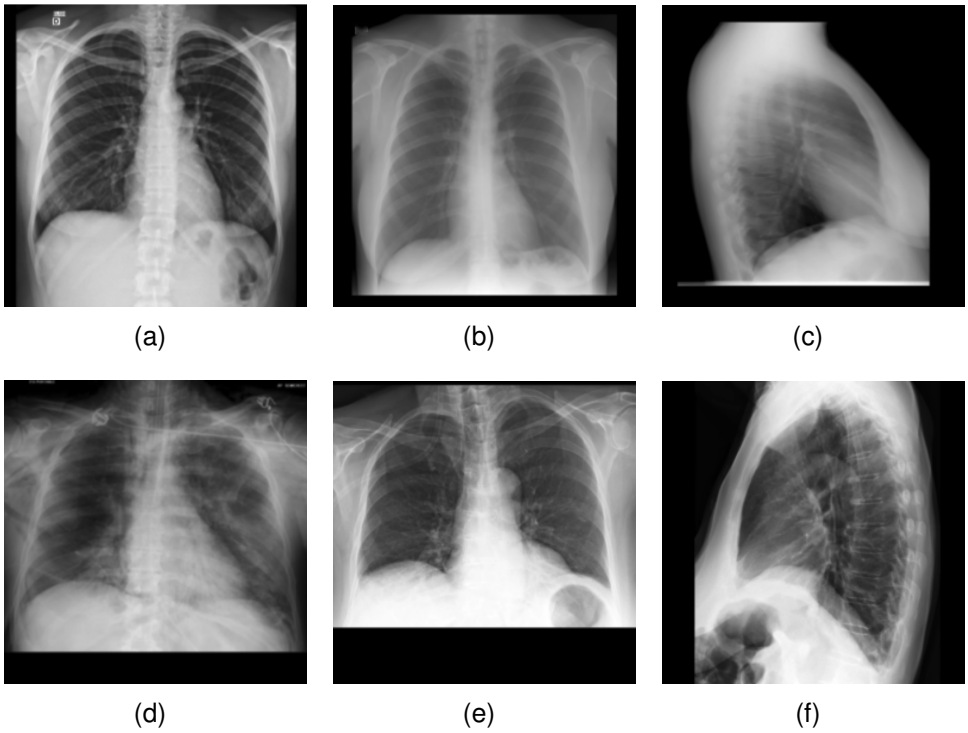
4.1.2 Bases de dados públicas

Além da base proposta, para compor uma solução completa para os desafios apresentados nesta tese, também foram avaliadas quatro bases de dados públicas. As mesmas possuem características distintas e se originam de diferentes regiões do mundo. Esse fator acrescenta uma heterogeneidade ao desenvolvimento da metodologia proposta que colabora para a sua atuação em diferentes cenários.

As bases de dados abaixo foram utilizadas em etapas distintas no desenvolvimento da metodologia proposta, sendo divididos em triagem e classificação de anormalidades.

4.1.2.1 Triagem

Indiana Dataset: Proposta por (Demner-Fushman et al. 2016) em 2016 e coletada no Hospital Universitário de Indiana nos Estados Unidos, possui 3.996 laudos médicos



(a) (b) (c)

(d) (e) (f)

Figura 7 – Exemplos de imagens normal (a)-(c) e anormal (d)-(f) utilizadas na base de dados proposta.

associados a 8.121 imagens de incidência frontal e lateral. No trabalho de Yates et al. (Yates, Yates e Harvey 2018), foi utilizado a divisão entre as classes normal e anormal, sendo utilizadas apenas as normais, sendo replicado no mesmo formato para avaliação da metodologia proposta neste trabalho.

NIH-RSNA: A Sociedade Radiológica da América do Norte (*Radiological Society of North America* - RSNA) em 2018 apresentou um desafio para detecção de Pneumonia em imagens de raios X de tórax. Sendo assim, foram disponibilizadas 30 mil imagens de raios X de tórax pertencentes ao conjunto de dados NIH Chest X-rays com uma nova rotulação realizada por seis especialistas certificados. A base foi subdividida nas classes normal, anormal com opacidade pulmonar e sem opacidade pulmonar (Shih et al. 2019).

4.1.2.2 Classificação de anormalidades

NIH Chest X-rays 14: A base de dados proposta por (Wang et al. 2017) foi disponibilizada com 112.120 imagens de raios X de tórax frontais, pertencentes a 30.805 pacientes únicos. Os autores apresentaram um método baseado em Processamento de Linguagem Natural (PLN) para realizar a extração dos rótulos com acurácia superior a 90%.

CheXpert: A CheXpert (Irvin et al. 2019) foi publicada em 2019 com o objetivo prover um conjunto de dados rotulado em 14 classes com o total de 224.316 radiografias, pertencentes a 65.240 pacientes. Assim como foi apresentado na NIH Chest X-rays

14, foi apresentada uma nova metodologia automática para extração de rótulos em laudos médicos. Quando comparado com o método de rotulação da NIH, foram obtidos resultados superiores para todas as classes em comum entre as bases de dados.

Na base de dados CheXpert, as radiografias de tórax são rotuladas com base na presença ou ausência de várias condições médicas. Essas rotulações são feitas de forma automática e/ou manual, e são classificadas em três categorias principais:

- Positivo (1): Indica a presença da condição na imagem;
- Negativo (0): Indica a ausência da condição na imagem;
- Incerteza (U): Indica que há incerteza sobre a presença ou ausência da condição, geralmente devido à ambiguidade na imagem ou na interpretação clínica.

Dentre as principais abordagens para utilização dessa base de dados, temos a U-zeros, onde todas as anotações incertas (U) são tratadas como se fossem negativas (0). Isso significa que o modelo será treinado considerando que as incertezas representam a ausência da condição. Essa abordagem foi utilizada na avaliação da metodologia proposta neste trabalho.

Na Tabela 6, apresentamos um resumo das bases de dados que possuem múltiplas classes relacionadas a detecção de anormalidades, as bases Indiana Dataset e NIH-RSNA foram utilizadas no somente na triagem de exames normais ou anormais. Desconsiderando imagens que não apresentam achados ou patologias, foram contabilizadas 24 classes que denotam alterações ou patologias a serem identificadas nos raios X nas bases CheXpert e NIH Chest X-rays 14.

4.2 Metodologia Proposta

Neste estudo, propomos uma metodologia hierárquica para auxiliar no diagnóstico de exames de radiografia de tórax, como mostrado na Figura 8. A metodologia é subdividida em três etapas, a primeira consiste no pré-processamento da base de imagens, adequando as imagens para a entrada das arquiteturas propostas.

A segunda etapa envolve a triagem de exames com alta probabilidade de serem normais ou anormais utilizando comitê de CNNs. Para o seu desenvolvimento, avaliamos três arquiteturas de CNNs amplamente utilizadas para classificação de radiografias do tórax. Sendo assim, classificamos os exames de acordo com o fator de confiança e definimos três classes de respostas de acordo com a probabilidade combinada entre as arquiteturas avaliadas.

Tabela 6 – Resumo do dataset proposto e os principais datasets utilizados na literatura para a classificação de anormalidades.

Anormalidades	CheXpert	NIH Chest X-rays 14
Sem achados	22.381	60.361
Aparelhos de suporte	116.001	-
Fratura	9.040	-
Edema pulmonar	52.246	1.686
Consolidação	14.783	4.667
Pneumonia	6.039	1.431
Lesão	9.186	-
Atelectasia	33.376	11.559
Enfisema	-	2.516
Pneumotórax	19.448	5.302
Hérnia	-	227
Fibroses	-	1.686
Derrame pleural	86.187	
Espessamento Pleural	-	3.385
Outros problemas pleurais	3.523	-
Cardiomegalia	27.000	2.776
Efusão	-	13.317
Cardiomediastino aumentado	10.798	-
Nódulos	-	6.331
Massas	-	5.782
Opacidade pulmonar	105.581	-
Infiltração	-	19.894
Espessamento	-	-
Aorta	-	-
Total	**224.316**	**112.120**

A terceira e última etapa consiste na subsequente classificação dos exames considerados como anormais de alta confiança de acordo com as suas anormalidades, envolvendo uma abordagem *multilabel*.

4.2.1 Etapa 1 - Pré-processamento

Devido à alta dimensionalidade, diferentes resoluções das imagens de entrada e o token metálico com o resultado do exame médico original, implementamos uma tarefa de pré-processamento para identificar a região de interesse (*Region of interest* - ROI) e adaptar as imagens de entrada ao padrão de entrada das CNNs. A Figura 9 descreve a etapa de pré-processamento implementada.

A entrada padrão para CNNs são imagens quadradas. Além disso, o grande número de operações realizadas torna desafiador o processamento de imagens com grandes dimensões. Assim, com base em testes empíricos, avaliações com médicos especialistas e resultados da avaliação de diversas dimensões apresentadas em Tang et al. (Tang et al. 2020), definimos o tamanho de entrada em 256$\times$256 *pixels*. No entanto, ao redimensionar para esta dimensão distorce as regiões da imagem, pois há uma grande diferença entre sua altura e largura. Assim, identificamos a região de interesse: a área do pulmão, antes da etapa de redimensionamento, aplicamos um limiar de Otsu (Otsu 1979)

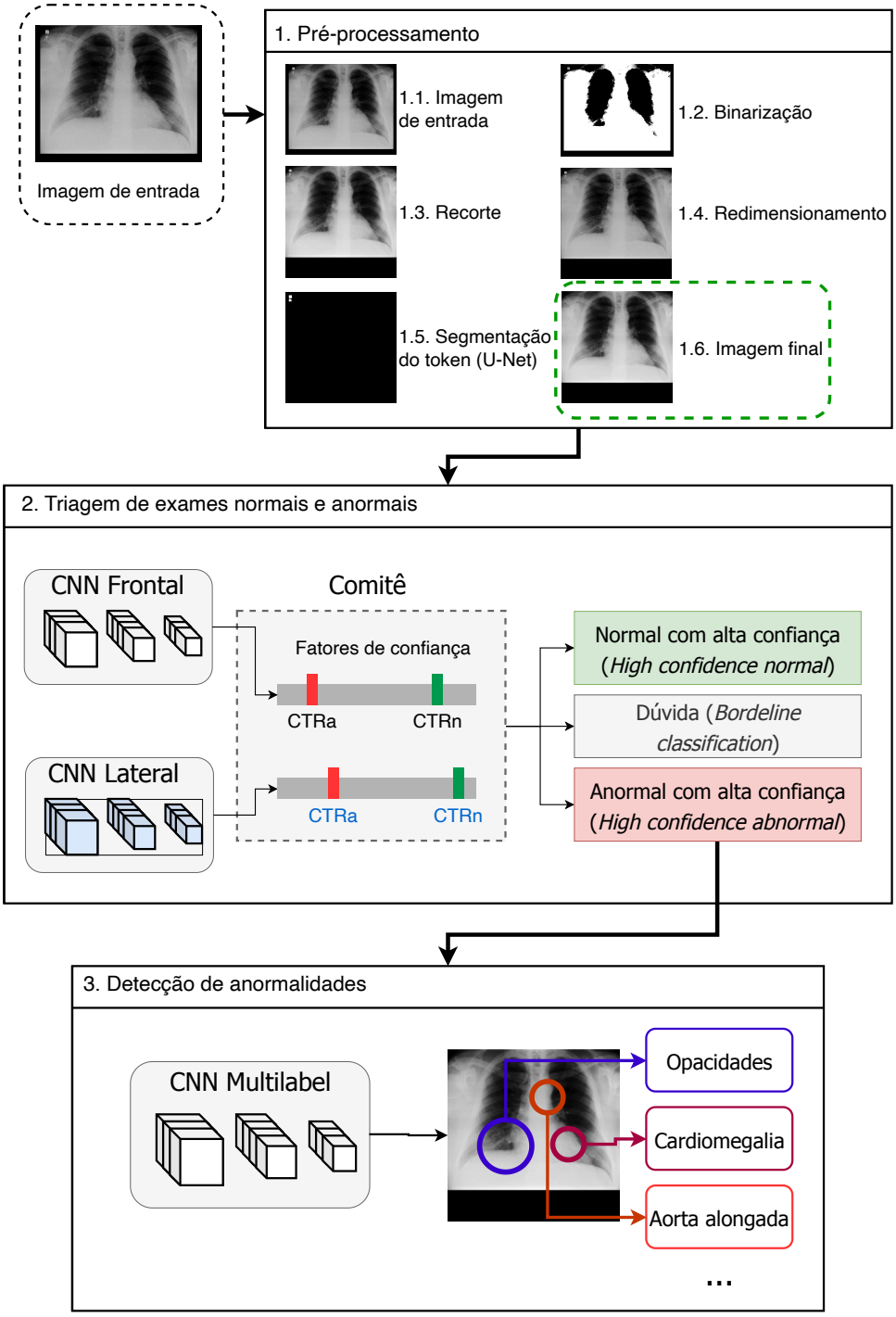


Figura 8 – Fluxograma da metodologia hierárquica proposta. Na etapa 1, apresentamos o pré-processamento da imagem de entrada. Na etapa 2, a triagem entre exames saudáveis e doentes, e por fim, na etapa 3 detectamos as principais anormalidades no exame doente.

para segmentar a região de interesse. Em seguida, o fundo da imagem é removido e a mesma é recortada para conter apenas a região do tórax. Após esse processo, é

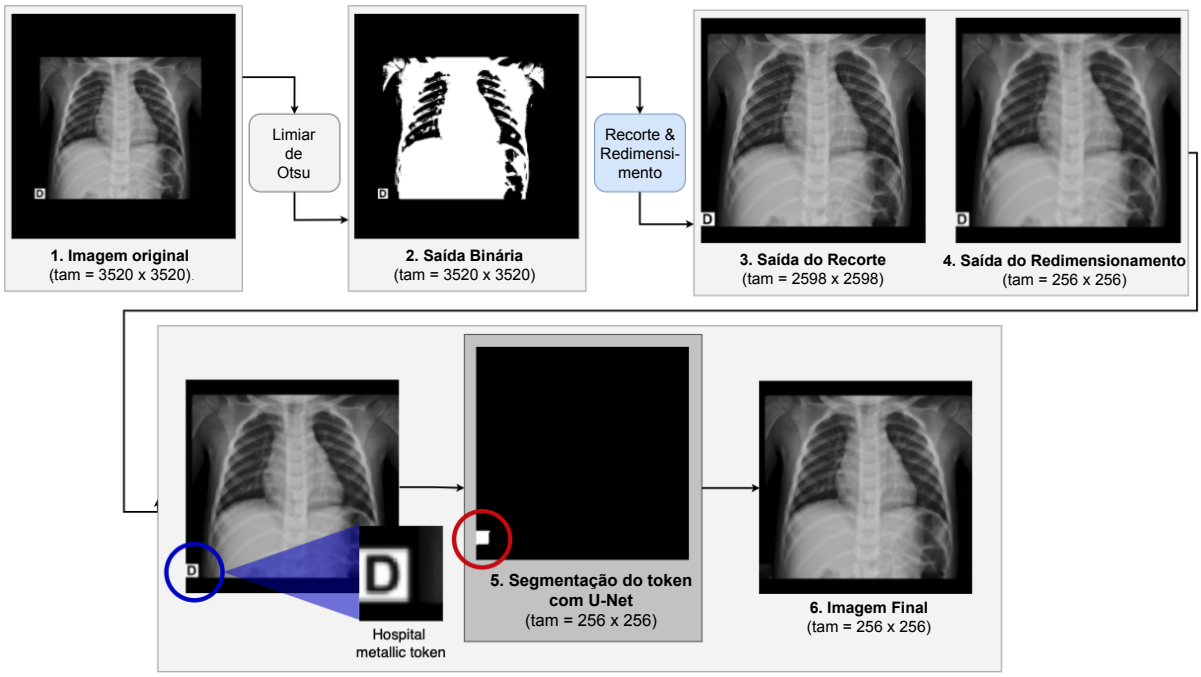


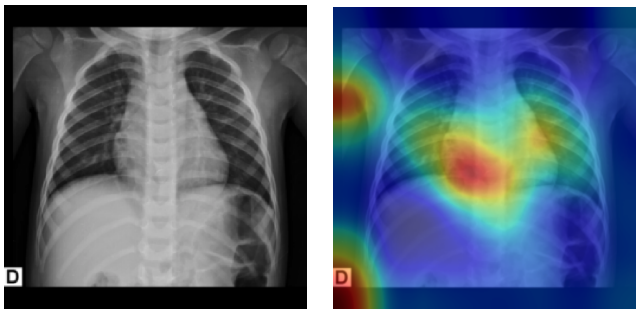
Figura 9 – Passos da etapa de pré-processamento proposta.

transformada em uma imagem quadrada usando *zero paddings*. Finalmente, a como resultado da transformação, a imagem finalmente é redimensionada novamente para para o tamanho desejado.

De acordo com Zech et al. 2018 e Geirhos et al. 2020, a presença de tokens metálicos em imagens de raios X pode influenciar o aprendizado de CNNs. A Figura 6 mostra a influência que um token metálico pode ter no processo de aprendizagem das CNNs. Assim, ele ilustra um mapa de ativação do VGG-16 para a incidência frontal em um cenário com e sem token. No cenário onde a imagem de entrada possui o token (Figura 10(a)), observa-se que uma das regiões onde o mapa de calor é mais intenso (vermelho) é a região do token (Figura 10(b)). Esse fato indica que a CNN considerou o token essencial para a tomada de decisões. Quando a imagem de entrada para a CNN é resultado de um processo de remoção de token (Figure 10(c)), o mapa de ativação (Figure 10(d)) na região onde o token foi removido é predominantemente azul, o que sugere que a região não é considerada relevante para a previsão.

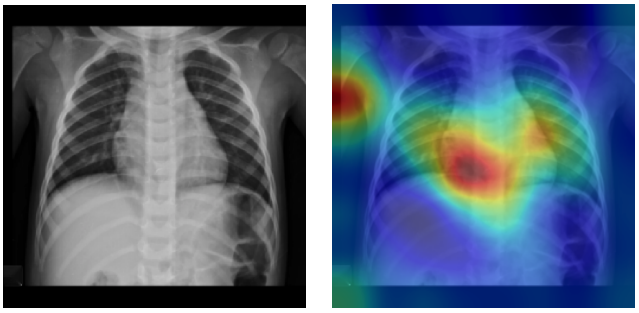
Dessa forma, a arquitetura de aprendizado profundo aprenderá o padrão de token e não os recursos críticos existentes no exame. Portanto, propomos a segmentação de tokens através da U-Net (Ronneberger, P.Fischer e Brox 2015). Escolhemos uma CNN para segmentar essa região porque não há padrões de tamanho e localização dos tokens nos exames. Assim, treinar uma U-Net com poucos filtros convolucionais e menos complexidade é mais eficaz para diferentes situações.

Para o treinamento e avaliação da U-net, o especialista rotulou manualmente o token de 239 imagens; então, usamos 80% desses dados para treinar a rede. Além disso,



(a) Imagem com token metálico

(b) Mapa de ativação para imagem com o token (a)



(c) Resultado do processo de remoção do token

(d) Mapa de ativação para imagem sem o token (c)

Figura 10 – Mapas de calor com as regiões que tiveram maior influência na previsão da CNN utilizada: após a remoção do token, a arquitetura da CNN utilizada não considera a região do token como uma região relevante para o resultado final.

usamos os 20% restantes para avaliar a segmentação das regiões do token. Esta metodologia obteve uma precisão de 99,96%. Em seguida, o token foi removido da imagem por meio da abordagem *Fast Marching Method* (FMM) proposto por Telea 2004, que considera as informações dos *pixels* vizinhos, partindo das bordas da região de interesse. Os *pixels* da região são substituídos pela soma ponderada e normalizada de todos os *pixels* da vizinhança.

O pré-processamento da imagem influencia diretamente na predição da CNN utilizada na metodologia proposta, as etapas de recorte e redimensionamento garantem o melhor aproveitamento da região de interesse, maximizando a extração das suas características, enquanto a remoção do token elimina qualquer viés na abordagem durante o treinamento. Na metodologia proposta, esta etapa garante a uniformidade das informações, pois foram obtidas de diferentes máquinas e com padrões distintos, auxiliando no aumento da generalização da metodologia proposta.

4.2.2 Etapa 2 - Triagem de exames normais e anormais

A triagem de exames é crucial para orientar o atendimento mais adequado ao paciente por parte do médico responsável. Identificar rapidamente exames sem alterações ou normais pode acelerar a identificação de outros problemas por meio de

outros exames. O médico em sua análise, aplica seu conhecimento tanto em imagens de incidência frontal quanto lateral, sendo assim, como primeira etapa da metodologia proposta, empregamos a utilização de um comitê de classificadores que irá analisar e combinar informações aprendidas nas duas incidências.

4.2.2.1 Comitê de classificadores proposto

Um comitê de classificadores consiste em combinar diferentes predições de diferentes modelos para emitir uma única resposta sobre os dados de entrada (Dasarathy e Sheela 1979). Dentre as principais vantagens do uso dos comitês, podemos citar a redução do *overfitting*, da variância e a minimização da instabilidade dos algoritmos de aprendizado.

A aplicação do comitê pode ser comparada com a avaliação feita por radiologistas na identificação de exames normais ou não. Essa característica comum se deve a uma ou mais incidências, principalmente, frontal e lateral, nas radiografias de tórax. A combinação da avaliação de cada incidência auxilia o médico na decisão final. Portanto, o comitê empregado neste trabalho atua de forma semelhante, pois duas CNNs são treinadas respectivamente com duas incidências de um exame. Assim, é possível produzir uma predição final usando regras predefinidas.

Com o intuito de fornecer respostas com alta probabilidade de serem normais ou anormais. Apresentamos o uso do Limiar de Confiança (*Confidence Threshold* - CTR) para definir as predições em Normal com Alta Confiança (*High Confidence Normal* - HCn), classificação Borderline (BC) ou Anormal com Alta Confiança (*High Confidence Abnormal* - HCa), de acordo com a probabilidade dada por cada CNN. O CTR auxilia na redução de erros e, consequentemente, no aumento da precisão da metodologia proposta. Dessa forma, a abordagem pode auxiliar na redução do esforço médico e contribuir para a melhoria da qualidade dos laudos médicos por meio da emissão de pré-laudos. Sendo assim, o médico apenas revisará a resposta dada pela metodologia. Podemos assim definir uma imagem como uma das três classes de acordo com:

$$R_i = \begin{cases} HC_n & \text{se } P_i > CTRn, \\ HC_a & \text{se } P_i < CTRa, \\ BC & \text{se } P_i > CTRa \text{ e } P_i < CTn, \end{cases} \tag{4.1}$$

onde R é a resposta obtida do comitê e P_i é a probabilidade do primeiro neurônio obtido pela camada de ativação softmax para uma imagem i. O fator de confiança representado por CTR pode assumir valores de acordo com a classe, sendo n normal e a anormal.

Vale ressaltar que os valores de CTRa e CTRn foram definidos empiricamente, variando de 0,51 a 1,0 para a classe normal e de 0,5 a 0 para a classe anormal. A

Figura 11 apresenta um exemplo de como o comitê construído funciona na classificação de imagens usando valores aleatórios para CTRa e CTRn para uma incidência frontal e uma incidência lateral.

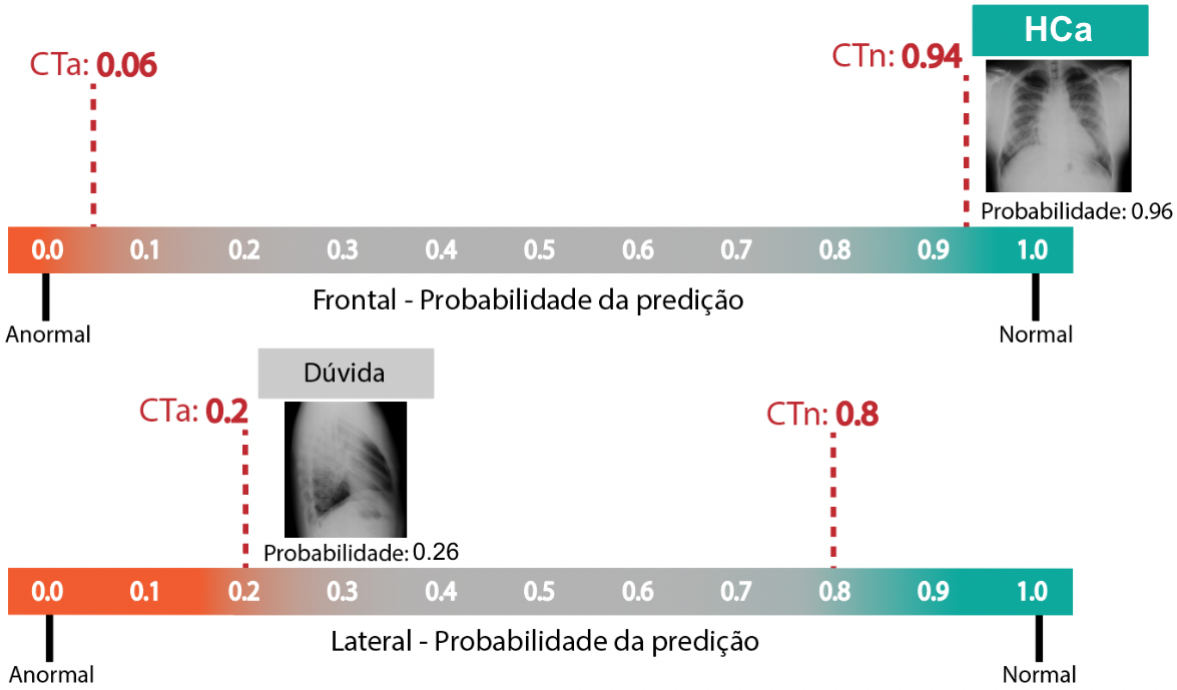


Figura 11 – Exemplo de como funciona a primeira etapa do comitê com fatores de confiança: A imagem foi classificada como normal na primeira linha, pois a probabilidade da predição (0,96) foi superior ao valor de CTRn (0,94). Na segunda linha, a imagem foi classificada como Borderline, pois a probabilidade de predição (0,16) ficou entre os valores de CTRa (0,2) e CTRn (0,8).

Após classificar todas as imagens do exame de acordo com os valores dos fatores de confiança, realizamos a segunda etapa do comitê, que decide a classe final. Da mesma forma que a avaliação por fator de confiança, seguimos o mesmo padrão de classes apresentado na decisão por incidência. No entanto, implementamos regras de acordo com o conhecimento médico para a tomada de decisão:

$$R_e = \begin{cases} HC_a & \text{se pelo menos um } R_i == HC_a, \\ HC_n & \text{se todos } R_i == HC_n \text{ ou o número de } HC_n > BC, \\ BC & \text{se todos } R_i == BC \text{ ou o número de } BC \geq HC_n, \end{cases} \tag{4.2}$$

onde e representa o exame avaliado. Nas regras implementadas, se houver pelo menos uma imagem i classificada como HCa, todo o exame será HCa. Portanto, para classificar o exame em HCn, a quantidade de imagens HCn deve ser maior que a quantidade de BC, ou todas são HCn.

4.2.3 Etapa 3 - Detecção de anormalidades

A detecção de anormalidades é crucial no entendimento de qual patologia aflige o paciente, acelerando o diagnóstico e mitigando qualquer prejuízo à vida do mesmo. Sendo assim, na metodologia proposta, empregamos a classificação *multilabel* ou multi-saídas, que é uma variação existente do problema multiclasses no cenário do aprendizado de máquina. Diferente do que encontramos nos multiclasses, onde o conjunto de dados possui n classes e para cada exemplo é atribuído um único rótulo, a classificação *multilabel* apresenta múltiplas classes não exclusivas, ou seja, um único exemplo pode possuir de 1 a n rótulos.

A representação proposta nesse tipo de classificação difere das demais pois os rótulos são apresentados como vetores, onde o seu tamanho será correspondente a $1 \times n$. Cada posição corresponde a uma classe do problema, sendo 0 ou 1 os valores que irão representar a ausência ou presença da classe.

Essa abordagem é comumente utilizada para a classificação de exames de raios X de tórax, uma vez que por meio dela é possível classificarmos múltiplas patologias em um único exame. Sendo assim, para compor a metodologia hierárquica, adotamos esse tipo de classificação, combinando suas características com as arquiteturas de CNNs do estado da arte.

A fase inicial da detecção de anormalidades consiste na aquisição da base de dados e seu tratamento. Optamos pelas bases CheXpert e NIH Chest X-rays 14, sendo duas das bases de dados mais utilizadas atualmente no estado da arte, possuindo 14 classes divididas de acordo com os principais achados encontrados nos raios X de tórax.

A segunda fase do processo consiste no ajuste fino de arquiteturas de CNNs. Dividimos os experimentos de acordo com as arquiteturas utilizadas na etapa 1 da metodologia proposta. Primeiramente as arquiteturas que obtiveram os melhores resultados na fase inicial, foram utilizadas como base inicial para o refinamento das bases de anormalidades. Depois avaliamos se o pré-treinamento na base de dados de triagem foi eficiente, utilizando os modelos pré-treinados na ImageNet. O uso de modelos pré-treinados em uma base de dados com imagens de mesma características auxilia na convergência da sua função de perda e no aumento da acurácia, que tendem a serem mais rápidas durante as épocas, auxiliando na obtenção de melhores resultados e acelerando a velocidade do treinamento.

4.3 Considerações Finais

A metodologia proposta é composta por três etapas principais, são elas: pré-processamento, triagem de exames e classificação de anormalidades. Cada etapa possui

técnicas e ferramentas próprias que se conectam e auxiliam na construção de um sistema CAD eficaz e robusto. A etapa de pré-processamento prepara a imagem de entrada que será utilizada pelas demais etapas. Já a etapa de triagem de exames consiste em diferençar exames normais e anormais com alta confiabilidade, dando suporte direto ao médico com a sua alta precisão e combinação de aprendizados por meio do comitê de classificadores. Já a terceira e última etapa, consiste na classificação das diferentes patologias encontradas em exames HCa (alta confiança de serem anormais).

5 Resultados e Discussão

O conjunto de experimentos para o desenvolvimento da metodologia proposta envolve duas principais etapas, são elas: triagem de exames normais e anormais e classificação de anormalidades. As mesmas possuem como principal foco, o uso de CNNs para a classificação das imagens e sua avaliação com métricas voltadas para classificação binária na triagem e multilabel na de anormalidades.

No processo de avaliação da metodologia proposta, consideramos a divisão do conjunto de dados em 80% para treino, 10% para validação e 10% para teste. Os dados foram separados por exame, ao invés de avaliarmos por paciente. Na avaliação da classificação por anormalidades, como foram empregadas base de dados públicas, a divisão foi realizada de acordo dos o estado da arte.

5.1 Triagem de exames normais e anormais

Para o desenvolvimento da etapa de triagem de exames da metodologia proposta, dividimos os experimentos em três fases, são elas: (1) seleção dos melhores modelos para incidência frontal, (2) seleção dos melhores modelos para incidência lateral e (3) construção da avaliação do comitê entre os modelos. Os critérios utilizados para a seleção dos modelos nas duas primeiras fases levaram em consideração métricas da literatura como acurácia, precisão, *recall*, especificidade, índice kappa e AUC. Para classificar os resultados, consideramos AUC como a métrica principal. Após a escolha dos modelos, o conjunto foi avaliado por meio dos fatores de confiança para as respostas com alta confiança de serem normais (HCn) e anormais (HCa). O critério de seleção das melhores combinações dos modelos do comitê foi o melhor resultado para as métricas de comprometimento médio da classe anormal (CMa) e normal (CMn).

Os resultados apresentados nesta seção foram obtidos sem o uso de aumento de dados. Apesar do aumento de dados ser uma técnica amplamente utilizada para melhorar o treinamento de CNNs e evitar o *overfitting*, não foi necessário neste trabalho. Realizamos testes com e sem aumento de dados nos experimentos iniciais e descobrimos que eles obtiveram desempenhos estatisticamente iguais. Sendo assim, optamos por não utilizar o aumento de dados, refletindo um cenário mais próximo do comumente encontrado pelos médicos. Além disso, verificamos que, devido ao número de imagens disponíveis, as CNNs com ajuste fino não sofrem com o problema de *overfitting*. A taxa de aprendizado empregada nos experimentos variou de 0,001 a 0,0001, o tamanho do *batch* utilizado é de 32.

5.1.1 Resultados para imagens de incidência frontal

A Tabela 7 apresenta os melhores resultados obtidos para as três arquiteturas avaliadas no conjunto de validação para as imagens de projeção frontal. As arquiteturas são apresentadas de acordo com as métricas obtidas e o número de camadas totalmente conectadas utilizadas. Além disso, ordenamos os resultados de acordo com a métrica AUC. Como segunda métrica para desempate, selecionamos o modelo com melhor desempenho no índice kappa. A Figura 12 apresenta a curva ROC e AUC para cada arquitetura indicada na Tabela 7. Para calcular a curva ROC e AUC, usamos a probabilidade de cada imagem pertencer à classe positiva e sua classe real. Essas informações foram extraídas da camada *softmax* das CNNs. O limiar utilizado para diferenciar as classes positivas e negativas foi de 0,5.

Tabela 7 – Cinco melhores resultados obtidos para imagens de incidência frontal usando o conjunto de validação com diferentes configurações de arquitetura (melhores resultados em negrito).

Arquitetura	**Camadas FC**	**Acc**	**P**	**R**	**S**	**K**	**AUC**
VGG-16	1024	**88,43%**	86,06%	81,40%	92,46%	**0,7472**	**0,8692**
VGG-16	1024 -512	87,99%	84,06%	**82,67%**	91,04%	0,7397	0,8684
VGG-16	1024 - 256	88,29%	87,14%	79,57%	93,28%	0,7423	0,8642
ResNet50	256	87,89%	**87,29%**	78,10%	**93,50%**	0,7324	0,8578
DenseNet121	512	87,67%	86,89%	77,85%	93,28%	0,7275	0,8555

Ainda, pela Tabela 7, podemos perceber que o melhor desempenho de validação foi alcançado pelo VGG-16 com 1024 neurônios na camada totalmente conectada, com 88,43% de precisão, 0,7472 quanto ao kappa, que é considerado muito bom e 0,8693 para AUC. Para o VGG-16 com duas camadas totalmente conectadas de 1024 e 512 neurônios, obtivemos uma AUC de 0,8685. No entanto, ao compararmos ambas, observamos uma sensibilidade superior para a VGG-16 1024, com 92,46%, comparado ao obtido pela VGG-16 1024-512. Dentre as arquiteturas com maior profundidade avaliadas, a ResNet50 obteve apenas 87,89% de acurácia, mas apresentou os maiores valores de precisão e especificidade.

Um fato a ser observado nos resultados obtidos é que com o aumento da quantidade de camadas totalmente conectadas, a VGG 1024-512 não obteve resultados superiores a sua configuração com apenas uma camada com 1024 neurônios. O mesmo pode ser observado para ResNet50, cujo melhor resultado foi obtido com apenas uma camada de 256 neurônios, e DenseNet121 com 512 neurônios.

Vale ressaltar que dentre as arquiteturas estudadas, a VGG-16, mesmo sendo mais antiga e rasa que as demais, obteve resultados superiores, ressaltando que foi a arquitetura que melhor generalizou as imagens no conjunto de validação para este conjunto de dados. Isso pode ser justificado devido a CNNs mais rasas produzirem

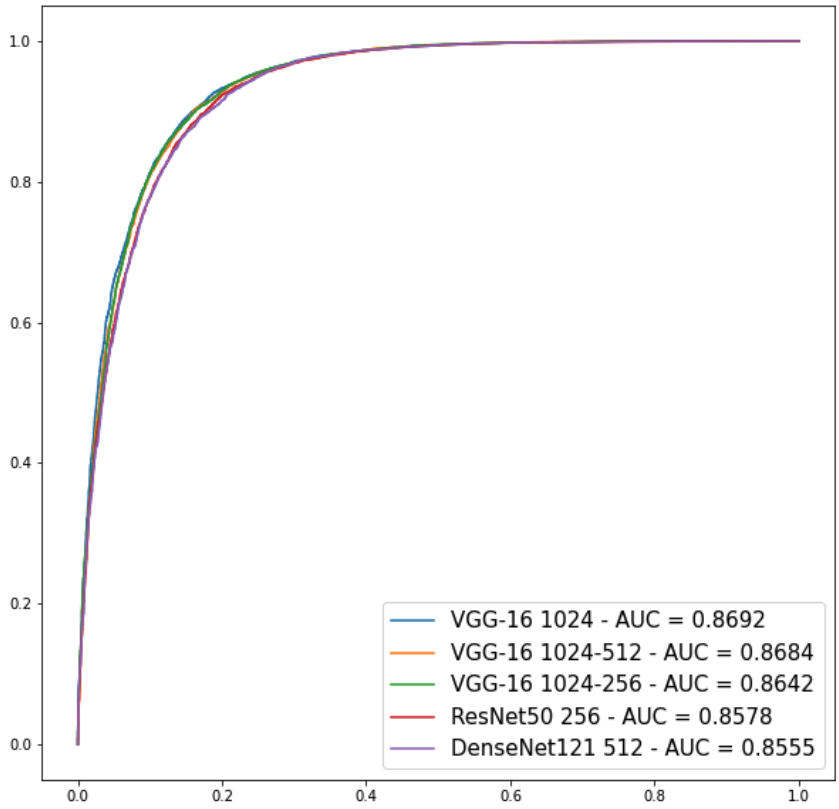


Figura 12 – Curva ROC e AUC dos melhores resultados obtidos para as imagens de incidência frontal.

melhores resultados para problemas binários ou com poucas classes do que arquiteturas mais profundas. Isso ocorre porque arquiteturas mais profundas são mais suscetíveis a *overfitting* em problemas binários (Vogado et al. 2021).

A Tabela 8 ilustra os resultados de um estudo de ablação para verificar a relevância da remoção de fichas metálicas nas cinco melhores abordagens definidas na Tabela 7. A partir dos resultados, observamos que o uso de tokens adicionou um viés no aprendizado da CNN e não apresentou resultados superiores em nenhuma das métricas ilustradas. Dentre as métricas avaliadas, apenas a acurácia apresentou resultados semelhantes com nível de significância de 5%. Foi utilizado o método t-Student para determinar se há uma diferença significativa entre os resultados.

Tabela 8 – Comparação entre os melhores resultados com e sem remoção de token em imagens de incidência frontal.

		Com Token			**Sem Token**		
Arquitetura	**Camadas FC**	**Acc**	**Kappa**	**AUC**	**Acc**	**Kappa**	**AUC**
VGG-16	1024	87,67%	0,7212	0,8445	**88,43%**	**0,7472**	**0,8692**
VGG-16	1024-512	87,84%	0,7171	0,8544	**87,99%**	**0,7397**	**0,8684**
VGG-16	1024-256	87,46%	0,7296	0,8433	**88,29%**	**0,7423**	**0,8642**
ResNet50	256	87,52%	0,7243	0,8543	**87,89%**	**0,7324**	**0,8578**
DenseNet121	512	87,07%	0,7158	0,8518	**87,67%**	**0,7275**	**0,8555**

5.1.2 Resultados para imagens de incidência lateral

A Tabela 9 apresenta os resultados obtidos pelos modelos para o conjunto de validação com imagens de incidência lateral. A Figura 13 apresenta a curva ROC e AUC para cada arquitetura indicada na Tabela 9.

Tabela 9 – Cinco melhores resultados obtidos para as imagens de incidência lateral usando o conjunto de validação com diferentes configurações de arquitetura (melhores resultados em negrito).

Arquitetura	**Camadas FC**	**Acc**	**P**	**R**	**S**	**K**	**AUC**
ResNet50	1024	83,62%	66,97%	**76,25%**	86,3%	0,5991	**0,8127**
VGG-16	1024 - 512	**85,0%**	**71,96%**	71,83%	89,8%	**0,6167**	0,8082
ResNet50	512	84,1%	68,88%	73,79%	87,86%	0,6028	0,8082
ResNet50	256	84,95%	71,77%	71,9%	89,7%	0,6157	0,808
VGG-16	1024 - 256	84,72%	71,75%	70,55%	**89,88%**	0,6075	0,8021

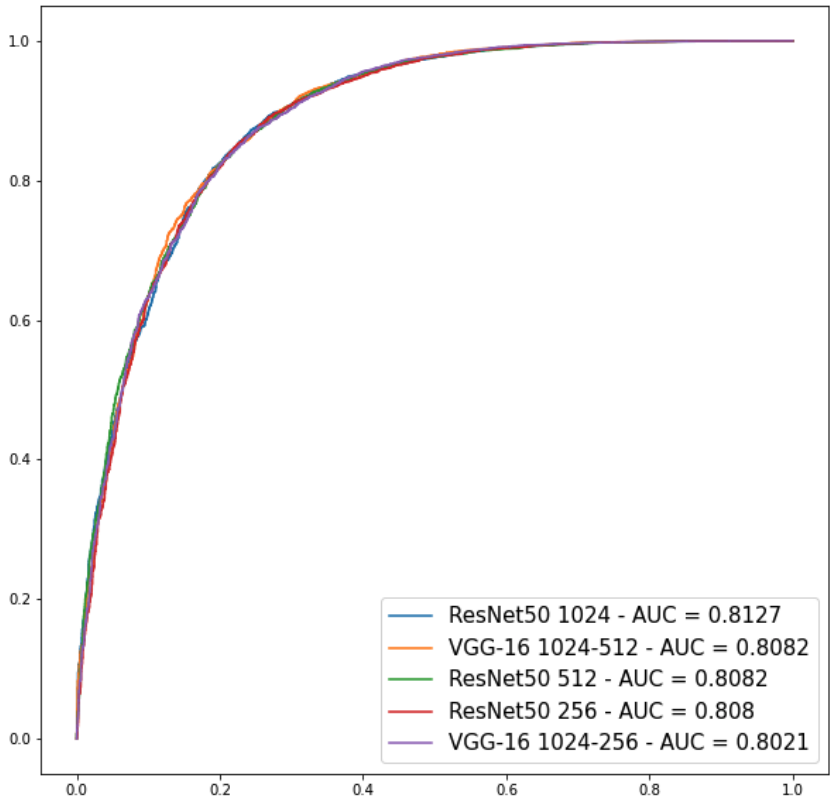


Figura 13 – Curva ROC e AUC dos melhores resultados obtidos para as imagens de incidência lateral.

Dos melhores resultados alcançados, destaca-se a ResNet50 com 1024 neurônios, que obteve 0,8127 de AUC e um recall de 76,25%. No entanto, mesmo com a melhor AUC, esta configuração de CNN teve menor acurácia, precisão, especificidade e kappa. Em contraste, o segundo melhor resultado obtido pela VGG-16 com 1024-512 alcançou os melhores resultados em três métricas, e o segundo melhor AUC empatou com a

ResNet50 com 512 neurônios. No entanto, a VGG-16 é considerada superior a ResNet50 de acordo com o critério de desempate definido pelo kappa obtido.

A Tabela 10 ilustra os resultados de um estudo de ablação para verificar a relevância da remoção do token metálico nas cinco melhores abordagens definidas na Tabela 9. Conforme discutido na incidência frontal, os resultados obtidos com a presença do token são inferiores aos obtidos sem a presença. Para a incidência lateral, apenas na configuração ResNet50 com 1024 neurônios, a abordagem com token obteve maior acurácia com 84,1% contra 83,62%. Porém, analisando o valor do kappa e principalmente da AUC, verificamos que a abordagem com a remoção foi superior.

Tabela 10 – Comparação entre os melhores resultados com e sem remoção de token em imagens de incidência laterais.

		Com Token			**Sem Token**		
Arquitetura	**Camadas FC**	**Acc**	**K**	**AUC**	**Acc**	**K**	**AUC**
ResNet50	1024	**84,1%**	0,5981	0,8025	83,62%	**0,5991**	**0,8127**
VGG-16	1024-512	84,89%	0,6083	0,7999	**85,0%**	**0,6167**	**0,8082**
ResNet50	512	82,29%	0,5725	0,8032	**84,1%**	**0,6028**	**0,8082**
ResNet50	256	83,41%	0,5922	0,8079	**84,95%**	**0,6157**	**0,8080**
VGG-16	1024-256	79,95%	0,5345	0,7944	**84,72%**	**0,6075**	**0,8021**

5.1.3 Comitê para triagem de exames

A partir dos resultados obtidos no conjunto de validação e na seleção dos cinco melhores modelos de CNNs, avaliamos o comitê para a triagem dos exames. Devido ao número de combinações existentes entre todos os modelos, reduzimos o escopo apenas para os dois melhores resultados com imagens de incidência frontal, e os dois melhores com imagens de incidência lateral. Sendo assim, dos modelos selecionados, estudamos oito combinações. A Tabela 11 indica os resultados do conjunto para o conjunto de validação junto com as cinco melhores correspondências alcançadas.

Dentre os objetivos traçados para a construção do comitê de classificadores, um dos principais é proporcionar alta confiança, ou seja, alta probabilidade de realmente pertencer a uma classe. Portanto, aumentamos a precisão da classe e diminuímos o erro. Por outro lado, há redução no número de respostas emitidas, agora apresentadas como BC. Além disso, para desenvolver e selecionar as melhores combinações, foi necessário impor regras para limitar a quantidade de respostas normais e anormais. Observamos através de experimentos empíricos realizados no conjunto de validação que com taxas de HCa e HCn acima de 30%, os valores de FOR e FDR estariam acima de 5% na maioria dos casos. Para limitar o erro a ser obtido, estabelecemos HCa e HCn entre 20% e 30%, buscando sempre o melhor valor de FOR e FDR entre essas duas faixas de resposta.

Durante os testes de construção do classificador ensemble, variamos os fatores de

confiança CTRn e CTRa para os modelos frontal e lateral e utilizamos o CM como parâmetro para balancear o número de respostas emitidas por classe e o erro obtido. Assim, observamos que obtivemos os melhores fatores de confiança a partir dos melhores valores de CM.

Tabela 11 – Resultados dos comitês ranqueados considerando diferentes arquiteturas para imagens de incidência frontal e lateral no conjunto de validação (melhores resultados em negrito).

CTRn - CTRa												
Frontal CNN	**Lateral CNN**	**Frontal**	**Lateral**	**BC**	**Qtd. de respostas**	**HCn**	**FDR**	**HCa**	**FOR**	**CMn**	**CMa**	**ACM**
VGG-16 1024 VGG-16 1024-512	VGG-16 1024 -512	0,94 - 0,06	0,8 - 0,2	8959 (47,07%)	10074 (52,93%)	29,57% (5629)	2,58% (145)	23% (4445)	4,5% (200)	70,28	**67,36**	**69,112**
VGG-1024	ResNet50 1024 VGG-16 1024-512	0,95 - 0,16	0,81 - 0,13	8222 (43,2%)	10811 (56,8%)	29,91% (5692)	2,78% (158)	27% (5119)	5,92% (303)	**70,29**	67,2	69,054
VGG-16 1024 VGG-16 1024-512	ResNet50 1024	0,93 - 0,15	0,81 - 0,04	8659 (45,49%)	10374 (54,51%)	29,62% (5637)	2,79% (157)	25% (4737)	4,33% (205)	70,17	67,35	69,042
VGG-16 1024	VGG-16 1024-512	0,95 - 0,16	0,84 - 0,07	8385 (44,06%)	10648 (55,94%)	29,82% (5675)	2,91% (165)	26% (4973)	5,43% (270)	70,18	67,19	68,984
VGG-16 1024	ResNet50 1024	0,94 - 0,14	0,83 - 0,04	8510 (44,71%)	10523 (55,29%)	29,81% (5673)	3,0% (170)	25% (4850)	5,03% (244)	70,12	67,17	68,94

Dentre os resultados obtidos, verificamos que nas cinco combinações apresentadas, o número de respostas foi superior a 50%. Para as CNNs frontais, os valores de CTRn foram superiores a 0,9 de probabilidade, enquanto para as CNNs laterais, nenhum dos CTRn ultrapassou a faixa de 0,85. Acreditamos que os fatores de confiança das CNNs laterais foram menores porque alguns dos achados ou alterações não são visíveis nas imagens laterais. No entanto, não diminui o desempenho do sistema, pois, de acordo com o protocolo médico, apenas a visão frontal é necessária em alguns casos, não sendo necessário o exame com a visão lateral. Assim, para casos de exames apenas com incidência frontal, a probabilidade emitida pelo CNN frontal é a única considerada na classificação em HCa, HCn ou BC.

Ao analisar os valores de HCn, observamos que eles estavam mais próximos dos 30% definidos como limite, enquanto os valores de FDR eram, até então, equivalentes ou inferiores a 3%. Esse mesmo comportamento não foi observado na classe anormal, que apresentou valores de HCa distantes de 30% e taxas de FOR superiores a 5% em alguns casos. Esse fato demonstra que a capacidade dos modelos de classificar exames normais e que o desequilíbrio entre as classes, mesmo atenuado, interferiu no resultado final.

Inicialmente, para avaliação das arquiteturas, considerou-se a combinação de dois modelos, um frontal e outro lateral. No entanto, calculamos a média entre as probabilidades obtidas por dois modelos da mesma projeção para obter melhores resultados. Para classificar os resultados, calculamos a *Average Commitment Metric* (ACM) de acordo com os valores de CMn e CMa. Com o conjunto entre os modelos frontal e lateral, encontramos o melhor resultado utilizando os modelos frontal VGG-16 1024 e VGG-16 1024-512, e o modelo lateral VGG-16 1024-512. Além disso, observamos que, dentre os cinco melhores resultados, os três que obtiveram os maiores valores de ACM tiveram uma combinação de modelos.

Calculamos os resultados para o conjunto de teste a partir da melhor combinação de modelos e definição de valores de CTR. Na Tabela 12, pode-se observar que algumas métricas foram superiores ao conjunto de validação, demonstrando que a abordagem conseguiu generalizar os resultados para um conjunto nunca visto pelos modelos.

Tabela 12 – Resultados do conjunto considerando diferentes arquiteturas para imagens frontais e laterais no conjunto de dados de teste.

		CTRn - CTRa										
Frontal CNN	**Lateral CNN**	**Frontal**	**Lateral**	**BC**	**Qtd. de respostas**	**HCn**	**FDR**	**HCa**	**FOR**	**CMn**	**CMa**	**ACM**
VGG-16 1024 VGG-16 1024-512	VGG-16 1024 -512	0,94 - 0,06	0,8 - 0,2	8524 (45%)	10263 (55%)	32% (6010)	1,68% (101)	23% (4253)	4,91% (209)	71,78	66,10	68,97

Para o conjunto de dados de teste, os resultados para a classe normal foram superiores aos obtidos para o conjunto de validação, com HCn igual a 32% e FDR de 1,68%. Na classe anormal, os resultados obtidos foram próximos aos obtidos pelo conjunto de validação, com 23% de HCa e FOR de 4,91%. Vale ressaltar que os valores

de CTR foram definidos usando o conjunto de validação, portanto, são replicados para o conjunto de teste sem limitações quanto à quantidade de respostas HCa ou HCn. A metodologia proposta apresentou um número de respostas superior a 50% do total de casos conforme o conjunto de dados de validação.

Realizamos um teste Z (Congalton e Green 2008) para comparar e avaliar estatisticamente os resultados entre o conjunto proposto (Tabela 11) e as CNNs individuais, Frontal VGG-16 1024, Frontal VGG-16 1024 512 e Lateral VGG-16 1024-512 (Tabelas 7 e 9), com nível de significância de 5% para avaliar se os resultados foram significativamente diferentes entre si. Os resultados mostraram que a abordagem de conjunto alcançou um desempenho significativamente superior ao das CNNs individuais.

5.2 Classificação de anormalidades

Para o desenvolvimento da segunda etapa de classificação dos exames, visando uma metodologia completa para auxilio ao diagnóstico de raios X de tórax, utilizamos duas bases de dados públicas nessa avaliação, são elas: NIH Chest X-rays 14 e CheXpert, ambas com 14 classes. Na CheXpert, foram utilizadas apenas imagens de incidência Frontal. A base de dados proposta não foi empregada no processo de classificação de anormalidades por não apresentar um rotulação com alta confiança por parte de especialistas, como nas bases de dados. A utilização da mesma implicaria na obtenção de possíveis vieses durante o processo de treinamento e predição.

Os modelos utilizados nessa etapa tiveram sua camada de classificação alterada para a quantidade correspondente de classes de cada base de dados. Além disso, mantivemos a função de ativação como "binary"e a função de perda como "binary crossentropy". A AUC foi utilizada na definição dos melhores modelos, sendo calculada em cada época.

Na Tabela 13 apresentamos os resultados obtidos nas arquiteturas de CNNs utilizadas para a base NIH Chest X-rays 14. As três arquiteturas pré-treinadas na base de dados proposta, DenseNet, VGG-16 e ResNet obtiveram AUC média superior a 0,80.

O melhor resultado obtido na Tabela 13 foi apresentado na VGG-16 1024 com 0.8483. Quando comparado com os demais modelos, observamos que não somente a média foi superior, mas também a quantidade de patologias em que o método alcançou o melhor resultado, sendo 10 de 14. Vale destacar que resultados obtidos pela DenseNet121 512 na classificação de Fibrosis, Cardiomegalia e Hernia, sendo superiores aos demais modelos com uma grande margem de AUC. Para a classe Emphysema, o melhor resultado foi obtido pela ResNet50 256 que alcançou uma AUC de 0,9254.

Na Tabela 14, apresentamos as novas arquiteturas MobileNetV2 e NasNetMobile e seus resultados para a classificação de anormalidades na base NIH Chest X-ray 14.

Tabela 13 – Resultados dos cinco melhores modelos frontais pré-treinados com a base de dados da etapa de triagem de exames normais e anormais e aplicados ao ajuste-fino na base NIH Chest X-rays.

Classes	DenseNet121 512	VGG-16 1024 512	ResNet50 256	VGG-16 1024 256	VGG-16 1024
Atelectasia	0,8183	0,8054	0,7789	0,7970	**0,8433**
Consolidação	0,7841	0,7736	0,7534	0,7751	**0,7916**
Infiltração	0,7879	0,8034	0,7776	0,8129	**0,8251**
Pneumotórax	0,8684	0,8797	0,8642	0,8086	**0,8806**
Edema	0,8631	0,8309	0,7835	0,8339	**0,8701**
Enfisema	0,9077	0,9073	**0,9254**	0,7941	0,9054
Fibrose	**0,8650**	0,8147	0,8042	0,7538	0,8395
Efusão	0,8273	0,8446	0,8240	0,8513	**0,8773**
Pneumonia	0,7712	0,7621	0,7462	0,7519	**0,7740**
Espessamento Pleural	0,7876	0,7856	0,7745	0,7523	**0,8074**
Cardiomegalia	**0,9009**	0,8581	0,8472	0,8021	0,8867
Nódulo	0,8117	0,8171	0,7876	0,7992	**0,8359**
Massa	0,8437	0,8341	0,7893	0,8131	**0,8556**
Hérnia	**0,9306**	0,8271	0,8443	0,7319	0,8841
AUC Média	0,8405	0,8246	0,8072	0,7912	**0,8484**

Nesse cenário, ambos foram pré-treinados com a base binária contendo exames normais e anormais e posteriormente refinada. O melhor resultado foi obtido pela NasNetMobile com AUC média de 0,7071, contra 0,6572 da MobileNetV2. Ambas as arquiteturas são caracterizadas por serem compactas, proporcionando modelos leves e fáceis de implantar em um cenário hospitalar por meio da redução de custos. No entanto, ao compararmos com os resultados obtidos na Tabela 13, observamos a diferença no desempenho, onde a menor AUC obtida foi de 0,7912, superior a ambas as arquiteturas.

Tabela 14 – Resultados dos modelos MobileNetV2 e NasNetMobile pré-treinados com a base de dados da etapa de triagem de exames normais e anormais e aplicados ao ajuste-fino na base NIH Chest X-rays.

Classes	MobileNetV2	NasNetMobile
Atelectasia	0,6121	0,6745
Consolidação	0,6644	0,6697
Infiltração	0,6875	0,6995
Pneumotórax	0,7305	0,8008
Edema	0,7181	0,6798
Enfisema	0,7535	0,8028
Fibrose	0,5804	0,6907
Efusão	0,6742	0,7356
Pneumonia	0,6540	0,6743
Espessamento Pleural	0,6362	0,7251
Cardiomegalia	0,6458	0,7054
Nódulo	0,6496	0,7010
Massa	0,6446	0,7058
Hérnia	0,5504	0,6340
AUC Média	0,6572	0,7071

Por fim, para validar se o pré-treinamento na base de triagem é eficiente, realizamos quatro experimentos divididos em duas partes (Tabela 15). Inicialmente utilizamos os mesmos modelos das tabelas anteriores, mas com o pré-treinamento sendo realizado apenas ImageNet. Para isso, selecionamos as duas melhores arquiteturas da Tabela 13 (VGG-16 1024 e DenseNet121) e os modelos MobileNetV2 e NasNetMobile, que por sua vez não obtiveram os melhores resultados na etapa de triagem, mas

que devem ser avaliadas utilizando essa variação. O intuito desse experimento é constatar se o pré-treinamento na base de triagem proporcionou resultados superiores ao pré-treinamento apenas na ImageNet.

Tabela 15 – Resultados dos modelos com o pré-treinamento realizado apenas com a base Imagenet.

Classes	VGG-16 1024	DenseNet121 512	MobileNetV2	NasNetMobile
Atelectasia	0,8299	0,7837	0,7839	0,6093
Consolidação	0,7889	0,7601	0,7272	0,6459
Infiltração	0,8267	0,7897	0,7787	0,6393
Pneumotórax	0,8886	0,8422	0,8296	0,5921
Edema	0,8646	0,8265	0,7836	0,7313
Enfisema	0,8937	0,8995	0,8320	0,5412
Fibrose	0,8007	0,8169	0,7239	0,5474
Efusão	0,8707	0,7931	0,8079	0,6566
Pneumonia	0,7651	0,7496	0,7273	0,6233
Espessamento Pleural	0,7923	0,7647	0,7497	0,5678
Cardiomegalia	0,8862	0,8489	0,8038	0,6398
Nódulo	0,8285	0,8002	0,7255	0,5362
Massa	0,8502	0,8078	0,7157	0,5367
Hérnia	0,8557	0,9206	0,7532	0,5267
AUC Média	0,8387	0,8145	0,7673	0,5995

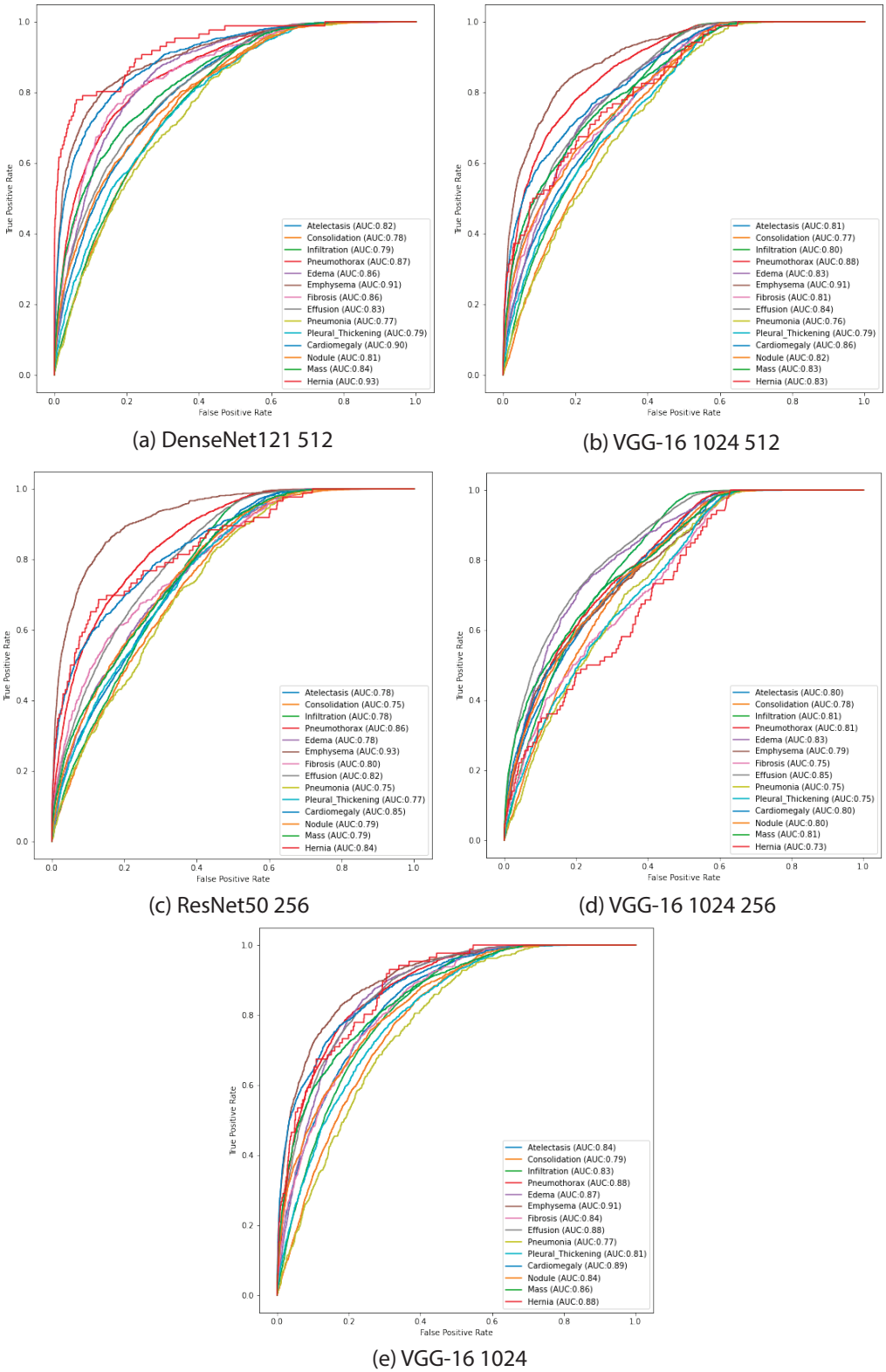
Comparando os resultados apresentados, os modelos VGG-16 1024, DenseNet121 512 e NasNetMobile obtiveram resultados superiores com o pré-treinamento realizado com a base de raios x da triagem. Para a VGG-16, alcançamos 0,8483 contra 0,8387 de AUC média. Na DenseNet121 o resultado foi de 0,8405 com triagem e 0,8145 com ImageNet e na NasNetMobile obtivemos 0,7071 contra 0,5995. A única arquitetura que apresentou resultados superiores com a ImageNet foi a MobileNetV2 com 0,7673 contra 0,6572 na triagem.

Na Figura 14, ilustramos os resultados obtidos também na Tabela 13. No entanto, incluímos a curva ROC, trazendo uma representação visual dos mesmos.

5.3 Discussão

A partir da Tabela 16 é possível comparar os resultados alcançados pelas metodologias propostas no estado da arte sobre a classificação binária dos exames em normais e anormais. Por outro lado, a partir da Tabela 17, é possível comparar os resultados de HCn obtidos na metodologia proposta com os de Dyer et al. 2021, único trabalho encontrado na literatura que, assim como a metodologia proposta, se propõe a emitir um diagnóstico inicial em exames com alta confiabilidade.

Na Tabela 16, são indicados os resultados obtidos para o conjunto de testes com a metodologia proposta e valores de CTR iguais a 0,5, que é o valor padrão encontrado na literatura para problemas binários. Esta metodologia obteve 87,54% de acurácia e 0,8721 quanto ao ROC, o que destaca a generalização dos modelos propostos para o conjunto de teste. Além disso, quando comparada com a metodologia da literatura, observamos



(a) DenseNet121 512

(b) VGG-16 1024 512

(c) ResNet50 256

(d) VGG-16 1024 256

(e) VGG-16 1024

Figura 14 – Curva ROC e AUC dos resultados obtidos nos modelos pré-treinados com a base de dados da etapa de triagem de exames normais e anormais.

que os resultados alcançados pela metodologia proposta são promissores e comparáveis aos resultados apresentados no estado da arte.

Um fato observado durante a análise dos resultados foi o número de imagens e os

Tabela 16 – Comparação da metodologia proposta com o estado da arte na classificação binária de exames de radiografia de tórax.

Trabalho	**Qtd. de imagens**	**Acc**	**AUC**
Yates, Yates e Harvey 2018	53.149	94,6%	-
Dunnmon et al. 2019	216.431	91,0%	-
Ellis et al. 2020	7.000	82,0%	-
Wong et al. 2020	128.886	-	0,8210
Tang et al. 2020	141.617	94,64%	0,9824
Metodologia Proposta	352.460	87,54%	0,8721

diferentes conjuntos de dados utilizados nos trabalhos de estado da arte encontrados. Enquanto a abordagem proposta usou mais de 300.000 imagens para desenvolver a metodologia, alguns autores apresentaram metodologias apenas com imagens frontais. Apenas o trabalho de Ellis et al. 2020 fez uso das duas incidências em seus experimentos. No entanto, os autores avaliaram a metodologia proposta usando um conjunto de dados menor com 7.000 imagens.

Levando em consideração as respostas dadas pela metodologia proposta baseada em comitês, comparamos seus resultados com os obtidos pela metodologia de Dyer et al. 2021, Tabela 17. Analisando a porcentagem de respostas com HCn, a metodologia sugerida por Dyer et al. 2021 atingiu 15%, enquanto a metodologia proposta obteve 32%. O erro para a classe normal da metodologia proposta foi menor que o apresentado por Dyer et al. 2021.

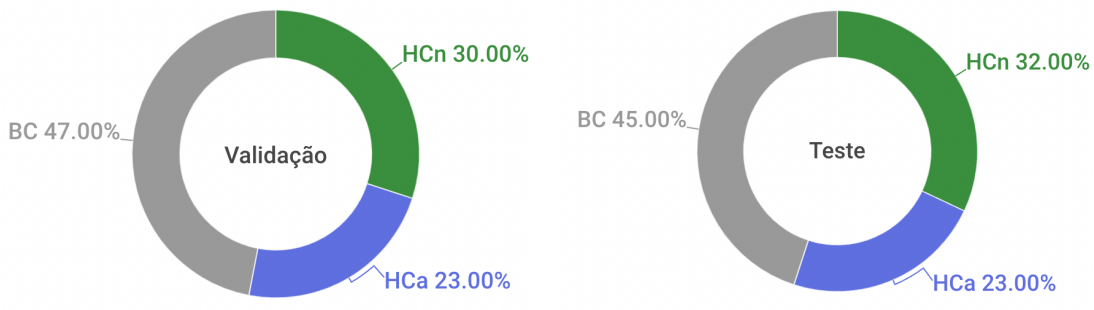
Tabela 17 – Comparação entre a metodologia proposta e a sugerida por Dyer et al. (Dyer et al. 2021).

Trabalho	**Qtd. de imagens**	**HCn**	**FDR**
Dyer et al. 2021	3.887	15%	2,3%
Metodologia Proposta	352.460	32%	1,68%

A Figura 15 ilustra a porcentagem de respostas HCn e HCa para os conjuntos de validação e teste. Observamos que a metodologia proposta classificou mais de 50% dos exames com precisão superior a 95% para ambas as classes.

Além da comparação indireta anterior, as Tabelas 18 e 19 apresentam comparações diretas com metodologias recentes do estado da arte.

Anteriormente, investigamos o desempenho da metodologia proposta nos conjuntos de imagens públicas usados por Yates, Yates e Harvey 2018 e Tang et al. 2020. O primeiro conjunto de imagens foi composto pelos bancos de dados Indiana Dataset e NIH, enquanto o segundo é uma versão simplificada do conjunto de dados NIH-RSNA. Nesta situação, não selecionamos os melhores parâmetros. Simplesmente executamos a metodologia proposta com os parâmetros previamente definidos com base no treinamento de nosso conjunto de dados. Os resultados obtidos são apresentados na Tabela 18, onde



(a) (b)

Figura 15 – Porcentagem de respostas com alta confiança para os conjuntos de validação (a) e teste (b) obtidos pela metodologia proposta.

Tabela 18 – Comparação entre os resultados obtidos pela metodologia proposta e os obtidos por metodologias do estado da arte em conjuntos de dados de imagens públicas (melhores resultados em negrito).

Trabalho	**Acc**	**P**	**R**	**S**	**AUC**
Indiana + NIH Datasets					
(Yates, Yates e Harvey 2018)	94,60%	**99,80%**	**94,60%**	93,4%	**0,98**
Metodologia Proposta	**98,85%**	62,00%	85,00%	**98,62%**	0,92
NIH-RSNA Dataset					
(Tang et al. 2020)	**92,34%**	88,09%	**97,40%**	87,55%	**0,9871**
Metodologia Proposta	89,63%	**92,83%**	86,23%	**93,21%**	0,8972

é possível observar que a metodologia proposta obteve resultados competitivos, sendo melhor em pelo menos duas métricas.

Além do experimento anterior, avaliamos o desempenho das metodologias do estado da arte em nosso conjunto de dados de imagens. Entre os autores que forneceram detalhes suficientes para reproduzir ou tornar o código fonte publicamente disponível, encontramos: Yates, Yates e Harvey 2018, Dunnmon et al. 2019 e Tang et al. 2020. Para esta avaliação, as implementações foram realizadas considerando todas as informações apresentadas nos respectivos trabalhos e foi utilizada apenas a incidência frontal visto que as abordagens foram desenvolvidas para serem aplicadas apenas nesta incidência. Além disso, os conjuntos de treinamento e teste foram os mesmos utilizados na definição da metodologia proposta.

A Tabela 19 apresenta os resultados desta avaliação. A metodologia de Tang et al. 2020 obteve acurácia de 87,19%, sendo considerada a melhor entre os três trabalhos encontrados na literatura. No entanto, o desempenho foi inferior ao obtido pela metodologia proposta, que obteve uma acurácia de 88,43%. De fato, as metodologias da literatura obtiveram um desempenho pior do que as cinco melhores arquiteturas avaliadas neste estudo usando incidência frontal (Tabela 7).

Na Tabela 20 apresentamos os principais resultados de autores do estado da arte

Tabela 19 – Comparação entre os resultados obtidos pela metodologia proposta e os obtidos pela metodologia do estado da arte no nosso conjunto de imagens (melhores resultados em negrito).

Trabalho	**Metodologia**	**Acc**	**AUC**
Yates, Yates e Harvey 2018	Ajuste fino com InceptionV3	86,89%	0,8364
Dunnmon et al. 2019	Ajuste fino com DenseNet121	86,95%	0,8456
Tang et al. 2020	Ajuste fino com ResNet18	87,19%	0,8450
Metodologia Proposta	mDFT VGG-16 1024	**88,43%**	**0,8693**

em comparação com a metodologia proposta neste trabalho. Com a VGG-16 1024, arquitetura que obteve o melhor resultado na triagem de exames normais e anormais, obtivemos a melhor AUC média dentre os trabalhos. Vale ressaltar que comparando os resultados individualmente por classe, o método proposto alcançou o melhor resultado em 5 das 14 classes, são elas: Atelectasia, Infiltração, Pneumonia, Nódulo e Massa. A justificativa para a obtenção de melhores resultados nessas classes se deve a base proposta utilizada no pré-treinamento da arquitetura, uma vez que a mesma possui uma heterogeneidade de anormalidades, mas foi obtida durante o período da COVID-19, ou seja, aumentando os casos de Pneumonia, Atelectasia e Infiltração. Além disso, na base anterior, combatemos também os erros ocasionados por nódulos, sendo assim, também contribuindo para o aumento na sua assertividade.

Os métodos de Nugroho 2021 e DSouza, Abidin e Wismüller 2019 também obtiveram excelentes resultados, com 0,8313 e 0,8268 de AUC média. Além disso, também apresentaram os melhores resultados em quatro das anormalidades cada um. Já Baltruschat et al. 2019 alcançou o melhor resultado apenas na classe de Hérnia.

Observamos que a metodologia proposta também alcançou resultados competitivos em todas as classes, demonstrando uma consistência nos resultados de todas as classes, com um desvio padrão de 0,038, enquanto DSouza, Abidin e Wismüller 2019 obteve 0,062 e Nugroho 2021 com 0,069 de desvio. Esse fato denota que além da melhor média, a proposta também consegue aprender bem diferentes características de cada classe e consequentemente generalizá-las.

Tabela 20 – Comparação entre os resultados obtidos pela metodologia proposta e os obtidos pela metodologia do estado da arte na base de dados NIH-14 (melhores resultados em negrito).

Anormalidade	Wang et al. 2017	Yao et al. 2018	Guendel et al. 2018	Baltruschat et al. 2019	Nugroho 2021	Sirazitdinov et al. 2019	DSouza, Abidin e Wismüller 2019	Mao et al. 2018	Metodologia
Atelectasia	0,7	0,733	0,767	0,763	0,7919	0,777	0,8143	0,7495	**0,8433**
Consolidação	0,703	0,711	0,745	0,749	0,7601	0,747	**0,811**	0,7283	0,7917
Infiltração	0,661	0,673	0,709	0,694	0,7051	0,694	0,7265	0,6869	**0,8251**
Pneumotórax	0,799	0,805	0,846	0,84	**0,8909**	0,845	0,8884	0,8451	0,8806
Edema	0,805	0,806	0,835	0,846	0,8609	0,838	**0,922**	0,834	0,8702
Enfisema	0,83	0,842	0,895	0,895	**0,9424**	0,906	0,9174	0,8699	0,9054
Fibrose	0,786	0,743	0,818	0,816	**0,8408**	0,822	0,8148	0,7978	0,8395
Efusão	0,759	0,806	0,828	0,822	0,8414	0,826	**0,8884**	0,8096	0,8773
Pneumonia	0,658	0,684	0,731	0,714	0,7366	0,731	0,7698	0,6954	**0,7740**
Espessamento Pleural	0,684	0,724	0,761	0,763	**0,808**	0,782	0,8076	0,7581	0,8075
Cardiomegalia	0,81	0,856	0,883	0,875	0,8917	0,887	**0,9129**	0,8687	0,8867
Nódulo	0,669	0,724	0,758	0,747	0,8036	0,768	0,7553	0,7255	**0,8360**
Massa	0,693	0,777	0,821	0,82	0,8356	0,825	0,8487	0,782	**0,8556**
Hérnia	0,872	0,775	0,896	**0,937**	0,9286	0,864	0,8388	0,8776	0,8842
AUC Média	0,745	0,761	0,807	0,806	0,8313	0,808	0,8268	0,7877	**0,8484**

Com o intuito de validar a metodologia desenvolvida com o uso da base de dados NIH Chest X-rays 14, selecionamos o melhor modelo (VGG-16 1024) pré-treinado na base de triagem e o ajustamos para a CheXpert. No entanto, no desafio proposto pelo *Stanford Machine Learning Group* foram avaliadas apenas cinco das quartoze classes, são elas: Atelectasia, Cardiomegalia, Consolidação, Edema e Efusão. Essas cinco classes foram selecionadas por serem algumas das condições mais comumente observadas em raios x de tórax e também por serem clinicamente significativas. A escolha dessas patologias permite que os modelos desenvolvidos a partir da base de dados sejam empregados em uma ampla gama de cenários clínicos, onde as mesmas são frequentemente diagnosticadas. Além disso, a seleção dessas classes foi influenciada pela disponibilidade de anotações de alta qualidade apresentadas no conjunto de dados original.

Na Tabela 21 apresentamos os resultados obtidos na base CheXpert. Alcançando AUC média de 0,8736, observamos que a metodologia proposta demonstra resultados competitivos com a literatura. Quando comparado com Irvin et al. 2019, observamos que nas patologias de Atelectasia, Consolidação e Efusão, alcançamos resultados próximos, abrindo margem para futuros experimentos.

Tabela 21 – Resultados das metodologia proposta para classificação de anormalidades avaliada com a base de dados CheXpert.

Metodologias	Atelectasia	Cardiomegalia	Consolidação	Edema	Efusão	AUC Média
Proposta	0,801	0,822	0,930	0,895	0,920	0,8736
Chexpert (Irvin et al. 2019)	0,811	0,840	0,932	0,929	0,931	0,8886

5.3.1 Explicabilidade

Atualmente, as CNNs fornecem excelentes recursos de aprendizado e generalização. No entanto, devido à sua complexidade, não apresentam transparência do que foi aprendido. Portanto, um aspecto crucial para entender os recursos mais relevantes usados na predição é a interpretação visual dos dados.

A interpretação visual pode ser classificada em duas categorias: interpretação de uma instância e interpretação geral da rede. A primeira categoria é dividida em métodos baseados em gradiente e métodos baseados em perturbação. Os baseados em gradiente, como *Class Activation Mapping* (CAM) e Grad-CAM (Selvaraju et al. 2017), utilizam as últimas camadas convolucionais para fornecer uma interpretação visual no nível de pixel e têm capacidade de discriminação de classe. Por outro lado, os métodos baseados na pertubação consideram um elemento como essencial para a tomada de decisão se a sua remoção alterar consideravelmente a saída. A importância deste elemento perturbador pode ser estimada comparando a saída da rede com e sem o elemento. Em imagens, por exemplo, é intuitivo que mudanças nos *pixels* que mais contribuem para um resultado levem a uma predição diferente. O método *Local Interpretable Model-agnostic*

Explanations (LIME) (Ribeiro, Singh e Guestrin 2016) é baseado em perturbação e representa a importância discriminativa da classe usando *superpixels*.

Para auxiliar na interpretação dos resultados obtidos pela metodologia proposta, implementamos e avaliamos exemplos classificados corretamente e incorretamente como normais e anormais. Além disso, apresentamos também o *ground-truth* com marcação de cores definida por um radiologista para diferentes achados ou patologias. Usamos Grad-CAM e LIME para fornecer uma interpretação visual dos exames. O Grad-CAM foi representado empregando o mapa de calor, sendo a região mais intensa a que mais contribuiu para a predição. O LIME apresenta os 10 principais *superpixels* que contribuíram positivamente (representados em verde) e negativamente (representados em vermelho) para a predição. Além disso, utilizamos as duas melhores arquiteturas frontais obtidas durante o desenvolvimento da metodologia e que fazem parte do comitê proposto.

As imagens foram selecionadas aleatoriamente para esta análise, e exemplos de imagens Verdadeiros Negativos e Verdadeiros Positivos são mostrados na Figura 16 e Figura 17, respectivamente.

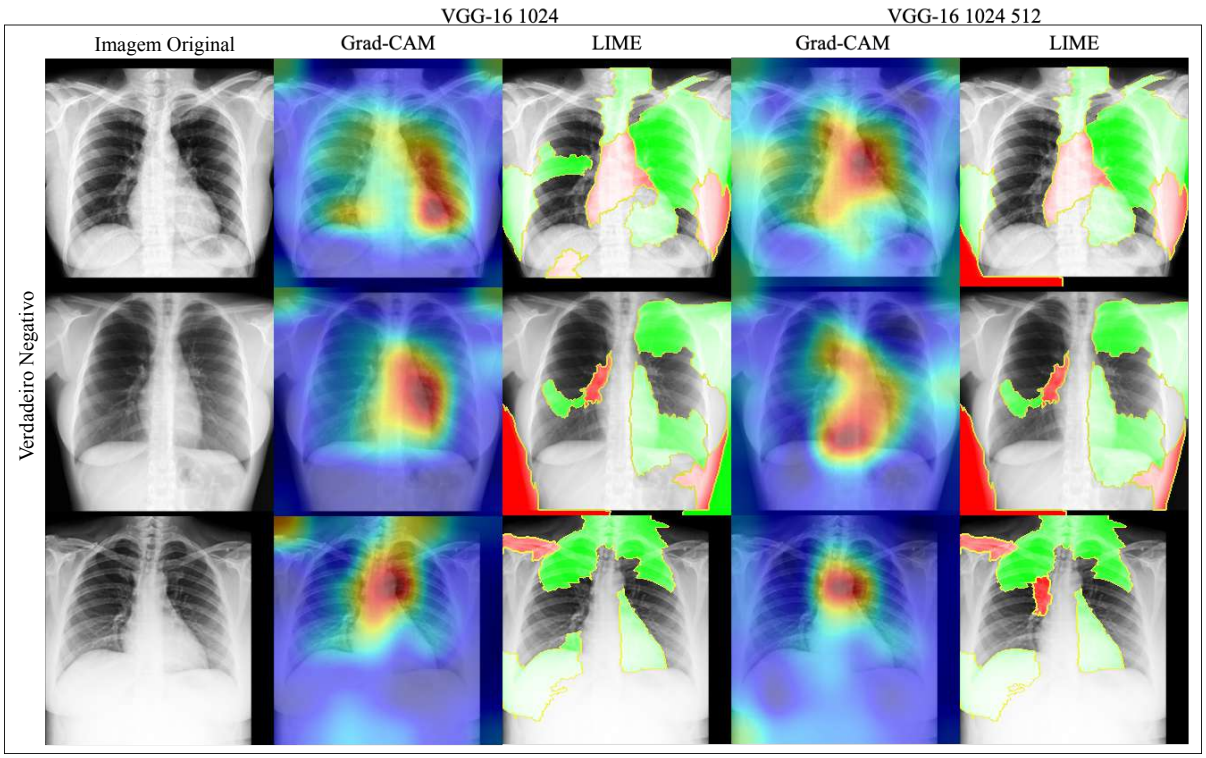


Figura 16 – Amostras com interpretações visuais para imagens corretamente classificadas como normais.

Nos exames da Figura 16, pode-se observar que, em ambas as visualizações, os modelos consideraram as regiões pulmonar e cardíaca como áreas críticas. Ainda, no terceiro exemplo, nota-se que a região de interesse foi a aorta, onde se concentrava a maioria dos casos de alterações. Concluímos que esta é uma região de grande relevância para classificação como HCa.

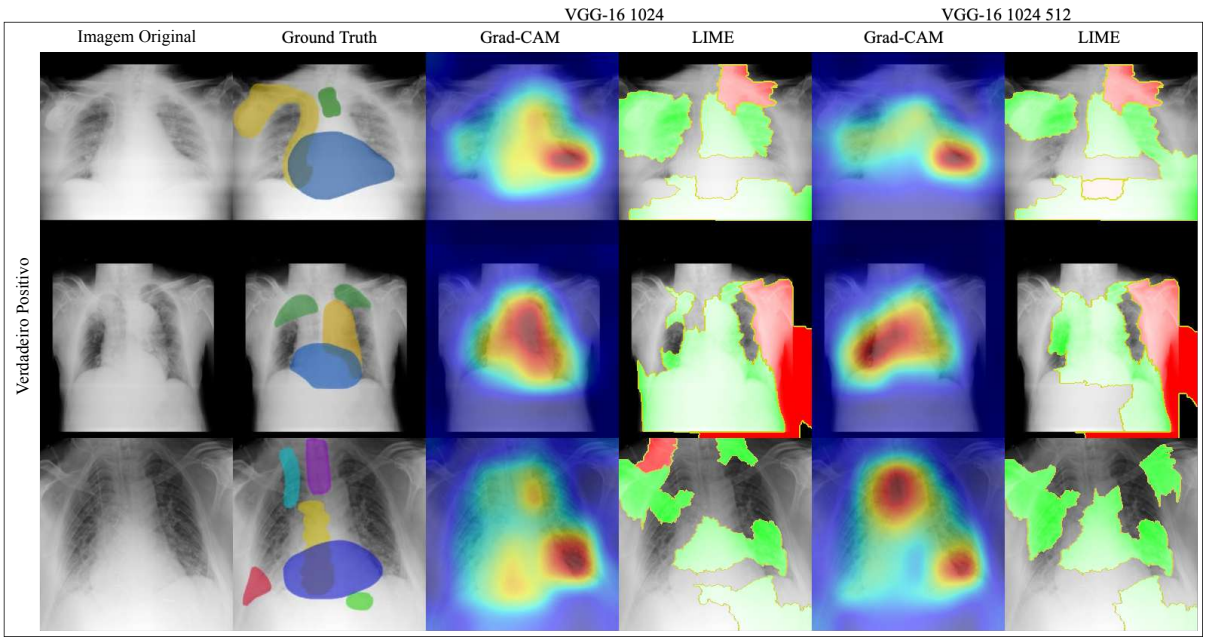


Figura 17 – Amostras com interpretações visuais para imagens corretamente classificadas como anormais.

Nos exemplos da Figura 17, observam-se três casos com alterações muito evidentes quando comparados com os exemplos Verdadeiros Negativos. O primeiro exame contém uma esternorrafia e marcapasso cardíaco e um aumento na área do coração. Pela interpretação, observamos que o Grad-CAM destacou as regiões do coração, denotando que a cardiomegalia foi a principal influência na predição como anormal. O LIME também destacou a área do marcapasso. No segundo exemplo, os pontos de interesse destacados escaparam da anormalidade, concentrando-se em regiões próximas ao coração que se configuravam como espessamento e aorta alterada. O terceiro exemplo mostra opacidades em ambos os pulmões e transparência reduzida da base do pulmão direito. Além disso, tubos, acesso central e clipes são visíveis. Ambos os métodos destacaram as regiões dos pulmões como as mais relevantes para a classificação. Além disso, achados adicionais foram considerados no Grad-CAM de VGG-16 1024 512.

Vale ressaltar que em todos os exemplos, o Grad-CAM gerou diferentes mapas de calor. Isso denota a capacidade dos comitês em utilizar características distintas da imagem para obter a predição final do exame, corroborando a robustez da metodologia proposta.

As Figuras 18 e 19 ilustram exames classificados incorretamente como anormais ou normais, ou seja, resultados Falso Positivo e Falso Negativo. Dentre os dois tipos de erros, destacamos maior gravidade nos FNs, pois podem apresentar riscos ao paciente. Vale ressaltar que, nesses casos, os exames tendem a se assemelhar visualmente a casos normais. Ainda assim, configuraram-se como anormais pela identificação de

alterações pouco perceptíveis nas imagens.

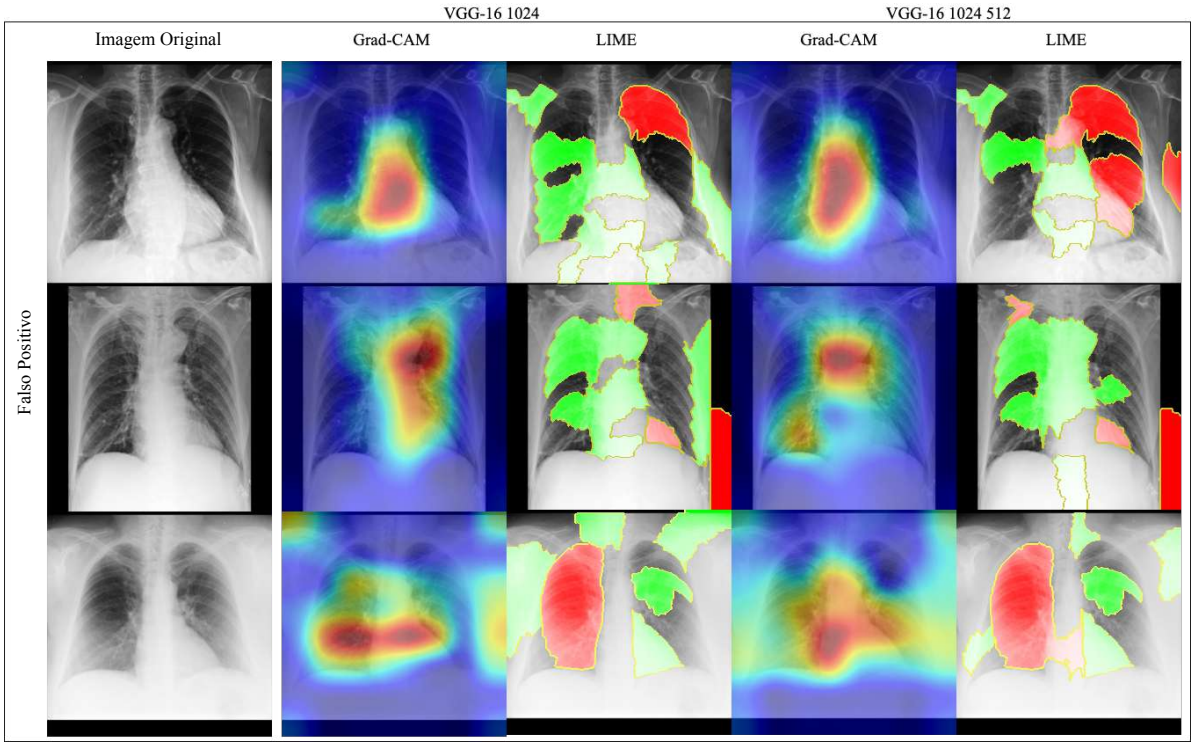


Figura 18 – Amostras com interpretações visuais para imagens classificadas incorretamente como anormais.

Para os casos FP, observamos resultados visualmente semelhantes aos mostrados na Figura 18 com a região pulmonar livre e alterações no volume cardíaco. Para prever os exames como normais, a região de interesse considerada foi a cardíaca. No primeiro exame apresentado, o achado visualizado foi o acesso venoso central destacado em amarelo. No segundo exame, o médico observou a artrose na coluna. No entanto, a alteração pode ser considerada invisível devido à redução na qualidade do exame devido ao redimensionamento. No terceiro exemplo, observou-se infiltração no pulmão direito utilizando o LIME com arquitetura VGG-16 1024. Entretanto, a rede considerou que esse achado contribuiu positivamente para a classificação normal.

Nos exemplos ilustrados na Figura 19, as interpretações fornecidas pelo Grad-CAM e LIME são focadas em regiões que apresentam riscos ao paciente, como coração e pulmões. Fatores que contribuem para o aparecimento desses erros são a qualidade da aquisição do exame, a interpretação do radiologista sobre o que deve ou não ser considerado uma alteração grave ou mesmo a necessidade de exames complementares para confirmação das alterações.

Na classificação de anormalidades, o uso de métodos de explicabilidade é crucial para auxiliar os médicos na identificação de patologias mais especificas. Nos casos em que temos múltiplas classes, sua relevância se torna maior, pois fornece um recurso visual preciso da região de interesse daquele achado. Com base nisso,

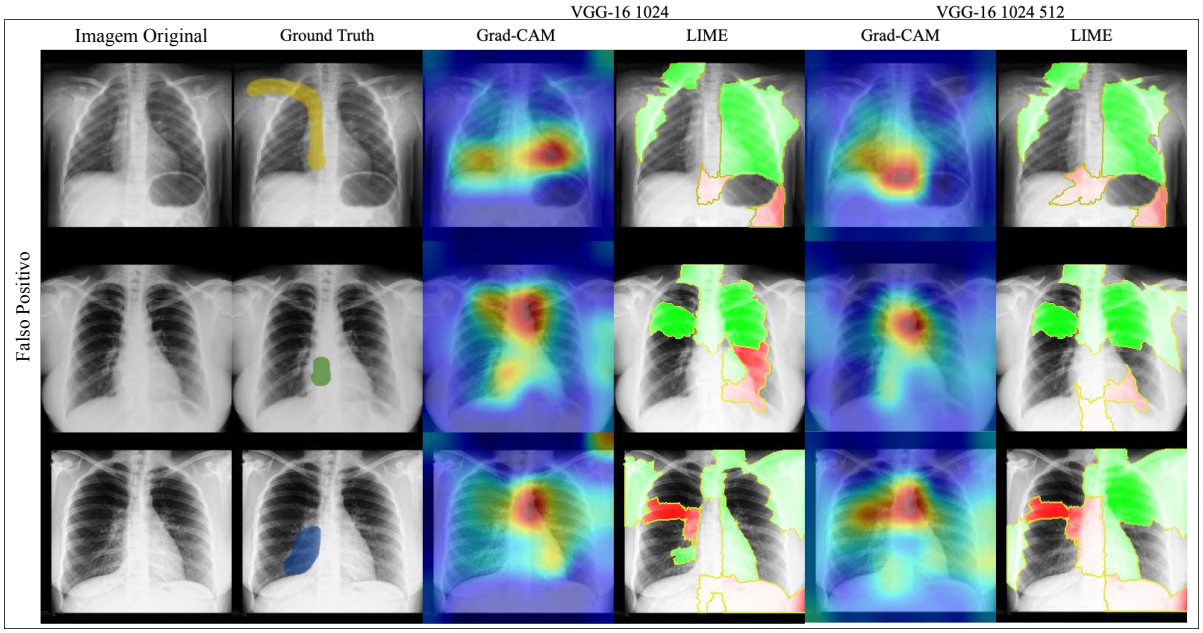


Figura 19 – Amostras com interpretações visuais para imagens classificadas incorretamente como normais.

selecionamos exemplos do conjunto de teste da base NIH-Chest X-ray 14 e apresentamos a representação obtida a partir do Grad-CAM na Figura 20.

Ainda na Figura 20, observamos a classificação correta de determinadas patologias dentro de um conjunto de até quatro anormalidades. O mesmo acontece para exames que possuem apenas uma única anormalidade e tivemos o acerto preciso da mesma.

5.3.2 Validação com médicos supervisores

A etapa de triagem de exames foi implementada em um servidor *Picture Archiving and Communication System* (PACs) para validar e certificar a qualidade das respostas com os radiologistas. Dois conjuntos de imagens estavam disponíveis, com 12.804 exames classificados como HCn e 7.183 classificados como HCa. A concordância obtida para a classe HCn foi de 98,60% (12.630) e para a FDR foi de 1,4% (174). Enquanto para a classe HCa, obtivemos 92,98% (6.679) de concordância e um FOR de 7,02% (502). A Tabela 22 apresenta os valores de FDR e FOR obtidos nos três cenários de avaliação: validação, teste e supervisão.

Tabela 22 – Valores FDR e FOR obtidos com diferentes conjuntos de avaliação.

Conjunto de avaliação	**FDR**	**FOR**
Validação	2,58%	4,50%
Teste	1,68%	4,91%
Supervisão	1,40%	7,02%

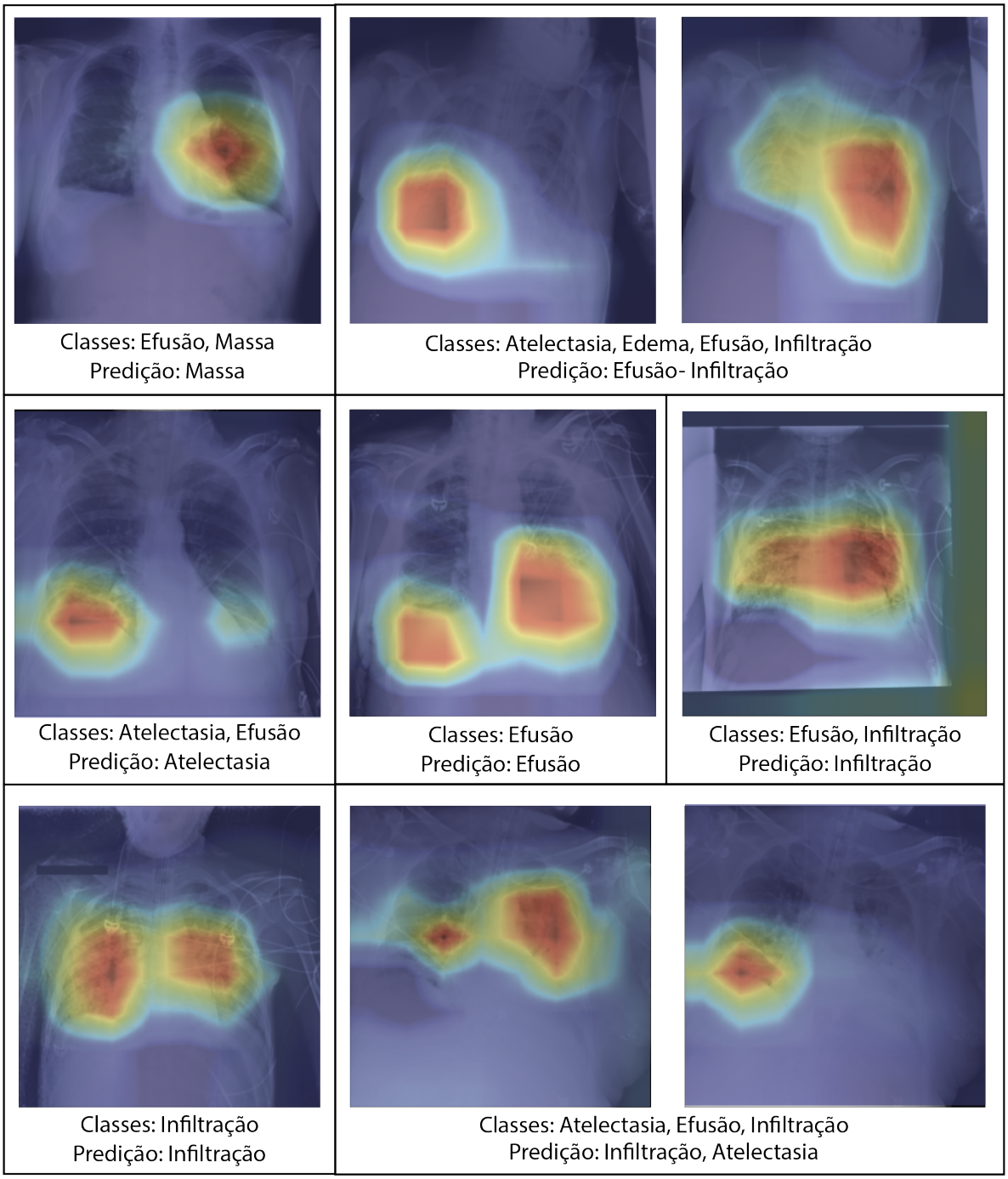


Figura 20 – Amostras com interpretações visuais com Grad-CAM para exames com múltiplas anormalidades na base de dados NIH-Chest X-ray 14.

Para exames normais, o valor FDR no conjunto de validação não correspondeu aos resultados dos conjuntos de teste e supervisão. O menor valor de FDR foi obtido na supervisão. O FOR obtido nos conjuntos de validação e teste foi semelhante, enquanto na supervisão obteve-se um valor maior, indicando mais Falsos Positivos.

Para exames anormais (métrica FOR), observamos que ainda são necessários aprimoramentos na metodologia proposta para aumentar a concordância com os

especialistas. Entretanto, podemos ressaltar que a interpretação dos médicos sobre o que deve ou não ser considerado uma patologia é fator determinante na acurácia de uma metodologia computacional. Outros fatores podem contribuir para a diferença nos resultados obtidos com os mesmos valores de CTRa e CTRn, como a diversidade existente no conjunto de dados com diferentes patologias.

Dentre os 172 exames classificados como anormais pelo especialista, grande parte contém granulomas (12,06%), que surgem devido ao processo de cicatrização de doenças prévias e geralmente não apresentam risco ao paciente. A Tabela 23 apresenta as cinco principais patologias ou achados entre os falsos negativos classificados pelo especialista na avaliação da supervisão.

Tabela 23 – As cinco principais patologias identificadas nos exames classificados incorretamente como normais na avaliação da supervisão.

Patologia	**Qtd. de exames**	**(% total de erros)**
Nódulos	23	13,21%
Granuloma	21	12,06%
Cardiomegalia	19	10,91%
Opacidades	16	9,19%
Consolidação	9	5,17%

Entre as patologias e achados apresentados na Tabela 23, nódulos e granulomas são difíceis de identificar por metodologias computacionais baseadas em classificação de imagem devido ao redimensionamento usual de imagem realizado. Além disso, as opacidades e consolidações, representando 9,19 e 5,17% do total de casos, respectivamente, em situações em que são mais discretas na base pulmonar, representam um desafio para a metodologia proposta.

5.3.3 Pontos fortes e limitações da metodologia proposta

Com base nos resultados apresentados, podemos observar que a metodologia proposta desenvolvida com base na definição de fatores de confiança é relevante neste contexto científico e aplicável em um cenário prático. Em comparação com metodologias encontradas no estado da arte, propusemos um pré-processamento mais robusto das imagens utilizadas no treinamento e não apenas um redimensionamento comum das imagens de entrada. Outra característica interessante é a utilização da incidência lateral no conjunto, que não é comumente levada em consideração na literatura. Vale ressaltar que essa incidência, quando presente no exame, é fundamental para a análise médica, pois diversas anormalidades como nódulos e granulomas não podem ser observadas adequadamente apenas com a incidência frontal.

Uma das limitações durante o processo de triagem é a classificação das imagens de entrada em apenas duas classes (binárias). Apesar disso, traz benefícios para médicos

e hospitais. Aos médicos pelo uso de fatores de confiança. Assim, consegue emitir respostas mais precisas, ajudando-os a reduzir o tempo de resposta para exames normais e dando um parecer prévio sobre a presença de patologias. Além disso, é possível utilizar respostas precisas no controle de filas para triagem de exames, otimizando o fluxo de atendimento. Além disso, vários hospitais estão isolados de grandes centros e não possuem radiologistas disponíveis para avaliar/laudar exames, portanto, com a metodologia proposta, é possível obter resultados com alto índice de acerto que auxiliam médicos residentes e médicos especialistas com poucos anos de experiência em seu diagnóstico. Adicionalmente, nota-se que o rastreio de exames saudáveis ou patológicos é fundamental para uma fase posterior onde será possível detectar patologias específicas.

Ressaltamos que os nódulos pulmonares (lesões sólidas e redondas) são um dos principais desafios enfrentados no desenvolvimento de metodologias computacionais para o diagnóstico de radiografias de tórax. Seguindo o fluxo da metodologia proposta, durante o redimensionamento da imagem, as características desse achado geralmente são perdidas, principalmente quando estão próximas à base pulmonar. Em alguns exames, o nódulo é a única alteração existente, e o restante do exame é totalmente saudável. Casos como esse aumentam o número de falsos negativos e exigem metodologias mais específicas para sua detecção. A Figura 21 mostra exemplos de radiografias de tórax com a presença de nódulos pulmonares.

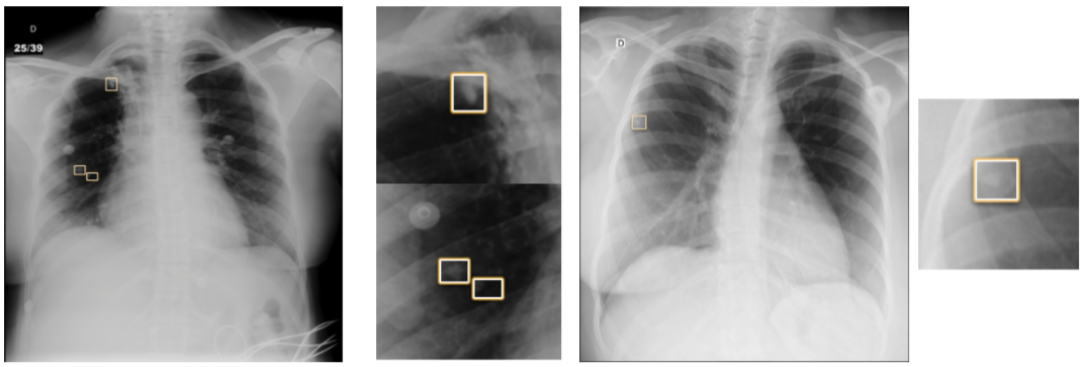


Figura 21 – Amostra de radiografias de tórax com presença de nódulos pulmonares.

6 Conclusão e trabalhos futuros

Este trabalho apresentou uma metodologia hierárquica baseada também em um comitê de classificadores de CNNs para predizer incidências de radiografia de tórax em normais ou anormais na fase de triagem e as anormalidades específicas de cada exame. Dentre os principais objetivos apresentados, conseguimos selecionar o melhor conjunto de CNNs e hiperparâmetros, utilizando técnicas de ablação que possibilitaram a obtenção de resultados que superaram diferentes métodos da literatura.

Dentre os principais desafios encontrados no desenvolvimento da metodologia, definir o conjunto de fatores de confiança para maximizar as respostas e reduzir o erro de cada classe foi o mais complexo. No entanto, por meio do desenvolvimento de um pré-processamento eficiente, em conjunto com a base de dados heterogênea e a seleção das melhores CNNs, alcançamos o resultado, validando o mesmo com conjuntos de teste e realizando uma auditoria com o médico especialista.

No trabalho proposto, ainda se encontram desafios, dentre eles: O conjunto de dados da etapa de triagem foi construído com exames eletivos, ou seja, patologias encontradas em casos de urgência não foram avaliados apropriadamente pela falta de exemplos, uma vez que dependemos de laudos para a classificação dos exames em normal e anormal. Na urgência os exames não são laudados. Além disso, existem limitações quanto ao próprio exame e a sua capacidade de detectar achados específicos, como nódulos e granulomas, sendo necessário complementar a abordagem proposta com um método específico para sua detecção.

Os resultados obtidos foram comparados com a literatura, onde demonstraram competitividade com os trabalhos do estado da arte tanto na etapa de triagem quanto na classificação de anormalidades. Alcançando 32% dos exames normais classificados com erro de 1,68% e 23% dos anormais com 4,91% de erro. Já na classificação de anormalidades, uma AUC média de 0,8484 para a base NIH e 0,8736 para a CheXpert. A metodologia proposta consegue assim atuar em duas etapas importantes no atendimento de pacientes, no processo de triagem, descartando pacientes com exames normais e para os casos mais graves, identificando as patologias, auxiliando em um tratamento específico e voltado para a doença em questão. Além disso, consegue atuar na geração de laudos normais para os casos eletivos, reduzindo o esforço do médico radiologista.

6.1 Trabalhos Futuros

Apesar dos bons resultados obtidos na triagem de exames, a metodologia ainda é suscetível a erros quanto a presença de nódulos, sendo necessário mais alguns

experimentos ou até mesmo a criação de métodos específicos para superar essa deficiência. Além disso, observamos que apesar da metodologia conseguir responder mais de 50% das amostras como HCn ou HCa, a quantidade de dúvidas ainda é alta, sendo necessário experimentos para aumentar o potencial de classificação da mesma. Observamos que durante a avaliação visual por meio da interpretabilidade, alguns exames apresentaram uma rotulação de caráter duvidosa, pondo em risco a classificação da metodologia. Seria interessante realizar uma segunda avaliação dessas imagens com um novo médico especialista para coletar seu parecer e realizar uma readequação na base de dados.

Já na etapa de anormalidades, apesar do excelente resultado na base NIH, na CheXpert a metodologia não foi tão eficiente. Com uma AUC média inferior, a metodologia proposta não alcançou bons resultados para a classificação de Edemas e Cardiomegalia, onde apresentaram maior deficiência. Novos experimentos são necessários para reavaliar o conjunto de dados da CheXpert com novos modelos de CNNs e um comitê para a etapa de classificação de anormalidades.

Além dos experimentos já citados, novas possibilidades surgem ao longo dos anos, dentre elas, a capacidade de gerar laudos de forma automática com uso de *Transformers* e Redes Neurais Generativas. Com esses experimentos, será possível emitir laudos ainda no processo de triagem de urgência dos pacientes, onde hoje atualmente oferece um alto custo para hospitais implementarem.

Referências

AKTAS, K.; IGNJATOVIC, V.; ILIC, D.; MARJANOVIC, M.; ANBARJAFARI, G. Deep convolutional neural networks for detection of abnormalities in chest x-rays trained on the very large dataset. **Signal, Image and Video Processing**, v. 20, p. 1–7, jul. 2022. Citado 2 vezes nas páginas 22 e 23.

ALLAOUZI, I.; AHMED, M. B. A novel approach for multi-label chest x-ray classification of common thorax diseases. **IEEE Access**, v. 7, p. 64279–64288, 2019. Citado na página 24.

BALTRUSCHAT, I. M.; NICKISCH, H.; GRASS, M.; KNOPP, T.; SAALBACH, A. Comparison of deep learning approaches for multi-label chest x-ray classification. **Scientific Reports**, v. 9, p. 6381, 2019. Citado 4 vezes nas páginas 24, 25, 64 e 65.

BEHZADI-KHORMOUJI, H.; ROSTAMI, H.; SALEHI, S.; DERAKHSHANDE-RISHEHRI, T.; MASOUMI, M.; SALEMI, S.; KESHAVARZ, A.; GHOLAMREZANEZHAD, A.; ASSADI, M.; BATOULI, A. Deep learning, reusable and problem-based architectures for detection of consolidation on chest x-ray images. **Computer Methods and Programs in Biomedicine**, v. 185, p. 105162, 2020. ISSN 0169-2607. Citado na página 21.

CHASSAGNON, G.; VAKALOPOULOU, M.; PARAGIOS, N.; REVEL, M.-P. Artificial intelligence applications for thoracic imaging. **European Journal of Radiology**, v. 123, p. 108774, 2020. ISSN 0720-048X. Citado na página 22.

CHEN, B.; LI, J.; GUO, X.; LU, G. Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays. **Biomedical Signal Processing and Control**, v. 53, p. 101554, 2019. ISSN 1746-8094. Citado na página 21.

CHONG, C. F.; WANG, Y.; NG, B.; LUO, W.; YANG, X. Image projective transformation rectification with synthetic data for smartphone-captured chest x-ray photos classification. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2210.05954>. Citado na página 24.

CHOPLIN, R. H.; BOEHME, J. M.; MAYNARD, C. D. Picture archiving and communication systems: an overview. **RadioGraphics**, v. 12, n. 1, p. 127–129, 1992. PMID: 1734458. Citado na página 18.

CONGALTON, R. G.; GREEN, K. **Assessing the Accuracy of Remotely Sensed Data: Principles and Practices**. 2. ed. Boca Raton: CRC Press, 2008. ISBN 9781420055122. Citado na página 58.

DASARATHY, B. V.; SHEELA, B. V. A composite classifier system design: Concepts and methodology. **Proceedings of the IEEE**, IEEE, v. 67, n. 5, p. 708–713, 1979. Citado na página 46.

DEMNER-FUSHMAN, D.; KOHLI, M. D.; ROSENMAN, M. B.; SHOOSHAN, S. E.; RODRIGUEZ, L.; ANTANI, S. K.; THOMA, G. R.; MCDONALD, C. J. Preparing a collection of radiology examinations for distribution and retrieval. **J. Am. Medical Informatics Assoc.**, v. 23, n. 2, p. 304–310, 2016. Disponível em: <https://doi.org/10.1093/jamia/ocv080>. Citado na página 39.

DMITRY, K. **Transfer learning method in the problem of binary classification of chest X-rays**. 2023. Disponível em: <https://arxiv.org/abs/2303.10601>. Citado 2 vezes nas páginas 22 e 23.

DSOUZA, A. M.; ABIDIN, A. Z.; WISMÜLLER, A. Automated identification of thoracic pathology from chest radiographs with enhanced training pipeline. In: MORI, K.; HAHN, H. K. (Ed.). **Medical Imaging 2019: Computer-Aided Diagnosis**. SPIE, 2019. v. 10950, p. 109503F. Disponível em: <https://doi.org/10.1117/12.2512600>. Citado 3 vezes nas páginas 24, 64 e 65.

DUNNMON, J. A.; YI, D.; LANGLOTZ, C. P.; Ré, C.; RUBIN, D. L.; LUNGREN, M. P. Assessment of convolutional neural networks for automated classification of chest radiographs. **Radiology**, v. 290, n. 2, p. 537–544, 2019. Citado 5 vezes nas páginas 21, 23, 62, 63 e 64.

DYER, T.; DILLARD, L.; HARRISON, M.; MORGAN, T. N.; TAPPOUNI, R.; MALIK, Q.; RASALINGHAM, S. Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm. **Clinical Radiology**, v. 76, p. 473.e9 – 473.e15, 01 2021. Citado 5 vezes nas páginas 11, 22, 23, 60 e 62.

ELLIS, R.; ELLESTAD, E.; ELICKER, B.; HOPE, M. D.; TOSUN, D. Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs. **Computers in Biology and Medicine**, v. 120, p. 103699, 2020. ISSN 0010-4825. Citado 3 vezes nas páginas 21, 23 e 62.

FUKUSHIMA, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. **Neural Networks**, v. 1, n. 2, p. 119 – 130, 1988. Citado na página 27.

GAO, F.; YOON, H.; WU, T.; CHU, X. A feature transfer enabled multi-task deep learning model on medical imaging. **Expert Systems with Applications**, v. 143, p. 112957, 2020. ISSN 0957-4174. Citado na página 17.

GEIRHOS, R.; JACOBSEN, J.-H.; MICHAELIS, C.; ZEMEL, R.; BRENDEL, W.; BETHGE, M.; WICHMANN, F. A. **Shortcut Learning in Deep Neural Networks**. 2020. Citado na página 44.

GOMES, J. C.; BARBOSA, V. A. de F.; SANTANA, M. A.; BANDEIRA, J.; VALENçA, M. J. S.; SOUZA, R. E. de; ISMAEL, A. M.; SANTOS, W. P. dos. Ikonos: an intelligent tool to support diagnosis of covid-19 by texture analysis of x-ray images. **Research on Biomedical Engineering**, p. 1–14, 2020. Citado na página 17.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning (adaptive computation and machine learning series). **Cambridge Massachusetts**, p. 321–359, 2017. Citado na página 28.

GUAN, Q.; HUANG, Y. Multi-label chest x-ray image classification via category-wise residual attention learning. **Pattern Recognition Letters**, v. 130, p. 259 – 266, 2020. ISSN 0167-8655. Image/Video Understanding and Analysis (IUVA). Citado na página 21.

GUENDEL, S.; GRBIC, S.; GEORGESCU, B.; ZHOU, K.; RITSCHL, L.; MEIER, A.; COMANICIU, D. **Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks**. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1803.04565>. Citado 3 vezes nas páginas 24, 25 e 65.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE Computer Society, 2016. p. 770–778. Citado 4 vezes nas páginas 9, 30, 31 e 33.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017. Disponível em: <https://arxiv.org/abs/1704.04861>. Citado 3 vezes nas páginas 9, 30 e 34.

Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. Densely connected convolutional networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 2261–2269. Citado 2 vezes nas páginas 30 e 31.

IRVIN, J.; RAJPURKAR, P.; KO, M.; YU, Y.; CIUREA-ILCUS, S.; CHUTE, C.; MARKLUND, H.; HAGHGOO, B.; BALL, R. L.; SHPANSKAYA, K. S.; SEEKINS, J.; MONG, D. A.; HALABI, S. S.; SANDBERG, J. K.; JONES, R.; LARSON, D. B.; LANGLOTZ, C. P.; PATEL, B. N.; LUNGREN, M. P.; NG, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. **CoRR**, abs/1901.07031, 2019. Disponível em: <http://arxiv.org/abs/1901.07031>. Citado 3 vezes nas páginas 24, 40 e 66.

ISMAEL, A. M.; ŞENGüR, A. The investigation of multiresolution approaches for chest x-ray image based covid-19 detection. **Health Information Science and Systems**, v. 8, p. 1–11, 2020. Citado na página 17.

ISMAEL, A. M.; ŞENGüR, A. Deep learning approaches for covid-19 detection based on chest x-ray images. **Expert Systems with Applications**, v. 164, p. 114054, 2021. ISSN 0957-4174. Citado na página 17.

ITANI, S.; LECRON, F.; FORTEMPS, P. Specifics of medical data mining for diagnosis aid: A survey. **Expert Systems with Applications**, v. 118, p. 300 – 314, 2019. ISSN 0957-4174. Citado na página 17.

KAMAL, U.; ZUNAED, M.; NIZAM, N. B.; HASAN, T. Anatomy-XNet: An anatomy aware convolutional neural network for thoracic disease classification in chest x-rays. **IEEE Journal of Biomedical and Health Informatics**, Institute of Electrical and Electronics Engineers (IEEE), v. 26, n. 11, p. 5518–5528, nov 2022. Citado 2 vezes nas páginas 24 e 25.

KARIM, F.; SHAH, M. A.; KHATTAK, H. A.; AMEER, Z.; SHOAIB, U.; RAUF, H. T.; AL-TURJMAN, F. Towards an effective model for lung disease classification: Using dense capsule nets for early classification of lung diseases. **Applied Soft Computing**, v. 124, p. 109077, 2022. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494622003702>. Citado 2 vezes nas páginas 24 e 25.

KATONA, T.; TóTH, G.; PETRó, M.; HARANGI, B. Developing new fully connected layers for convolutional neural networks with hyperparameter optimization for improved multi-label image classification. **Mathematics**, v. 12, n. 6, 2024. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/12/6/806>. Citado na página 25.

Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 27.

MAO, C.; YAO, L.; PAN, Y.; LUO, Y.; ZENG, Z. Deep generative classifiers for thoracic disease diagnosis with chest x-ray images. In: **2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. [S.l.: s.n.], 2018. p. 1209–1214. Citado 2 vezes nas páginas 24 e 65.

NUGROHO, B. A. An aggregate method for thorax diseases classification. **Scientific Reports**, v. 11, p. 3242, 2021. Citado 3 vezes nas páginas 24, 64 e 65.

OTSU, N. A Threshold Selection Method from Gray-level Histograms. **IEEE Transactions on Systems, Man and Cybernetics**, v. 9, n. 1, p. 62–66, 1979. Citado na página 42.

PHAM, H. H.; LE, T. T.; TRAN, D. Q.; NGO, D. T.; NGUYEN, H. Q. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. **Neurocomputing**, v. 437, p. 186–194, 2021. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231221000953>. Citado na página 24.

QIN, C.; YAO, D.; SHI, Y.; SONG, Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. **BioMedical Engineering OnLine**, v. 17, n. 113, p. 1 – 23, 2018. Citado na página 18.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016**. [S.l.: s.n.], 2016. p. 1135–1144. Citado na página 67.

RONNEBERGER, O.; P.FISCHER; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: **Medical Image Computing and Computer-Assisted Intervention (MICCAI)**. [S.l.]: Springer, 2015. (LNCS, v. 9351), p. 234–241. Citado na página 44.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision (IJCV)**, v. 115, n. 3, p. 211–252, 2015. Citado na página 30.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **2017 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 618–626. Citado na página 66.

SHIH, G.; WU, C.; HALABI, S.; KOHLI, M.; PREVEDELLO, L.; COOK, T.; SHARMA, A.; AMOROSA, J.; ARTEAGA, V.; GALPERIN-AIZENBERG, M.; GILL, R.; GODOY, M.; HOBBS, S.; JEUDY, J.; LAROIA, A.; SHAH, P.; VUMMIDI, D.; YADDANAPUDI, K.; STEIN, A. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. **Radiology: Artificial Intelligence**, Radiological Society of North America Inc., v. 1, n. 1, jan. 2019. ISSN 2638-6100. Citado na página 40.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **CoRR**, abs/1409.1556, 2014. Disponível em: <http://arxiv.org/abs/1409.1556>. Citado na página 30.

SIRAZITDINOV, I.; KHOLIAVCHENKO, M.; KULEEV, R.; IBRAGIMOV, B. Data augmentation for chest pathologies classification. In: **2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)**. [S.l.: s.n.], 2019. p. 1216–1219. Citado 2 vezes nas páginas 24 e 65.

TANG, Y.-X.; TANG, Y.-B.; PENG, Y.; YAN, K.; BAGHERI, M.; REDD, B. A.; BRANDON, C. J.; LU, Z.; HAN, M.; XIAO, J.; SUMMERS, R. M. Automated abnormality classification of chest radiographs using deep convolutional neural networks. **Digital Medicine**, IEEE, v. 3, n. 70, p. 1–8, 2020. Citado 6 vezes nas páginas 22, 23, 42, 62, 63 e 64.

TELEA, A. An image inpainting technique based on the fast marching method. **Journal of Graphics Tools**, Taylor and Francis, v. 9, n. 1, p. 23–34, 2004. Citado na página 45.

VOGADO, L.; VERAS, R.; AIRES, K.; ARAúJO, F.; SILVA, R.; PONTI, M.; TAVARES, J. M. R. S. Diagnosis of leukaemia in blood slides based on a fine-tuned and highly generalisable deep learning model. **Sensors**, v. 21, n. 9, 2021. ISSN 1424-8220. Citado 3 vezes nas páginas 27, 29 e 52.

WANG, X.; PENG, Y.; LU, L.; LU, Z.; BAGHERI, M.; SUMMERS, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, IEEE, Jul 2017. Citado 5 vezes nas páginas 21, 24, 25, 40 e 65.

WONG, K. C. L.; MORADI, M.; WU, J.; PILLAI, A.; SHARMA, A.; GUR, Y.; AHMAD, H.; CHOWDARY, M. S.; CHIRANJEEVI, J.; POLAKA, K. K. R.; WUNNAVA, V.; REDDY, D.; SYEDA-MAHMOOD, T. A robust network architecture to detect normal chest x-ray radiographs. In: **2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)**. [S.l.: s.n.], 2020. p. 1851–1855. Citado 3 vezes nas páginas 21, 23 e 62.

XU, X.; JIANG, X.; MA, C.; DU, P.; LI, X.; LV, S.; YU, L.; CHEN, Y.; SU, J.; LANG, G.; LI, Y.; ZHAO, H.; XU, K.; RUAN, L.; WU, W. **Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia**. 2020. Citado na página 17.

YANAS, J.; TRIANTAPHYLLOU, E. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. **Expert Systems with Applications**, v. 138, p. 112821, December 2019. Citado na página 17.

YAO, L.; PROSKY, J.; POBLENZ, E.; COVINGTON, B.; LYMAN, K. **Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions**. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1803.07703>. Citado 3 vezes nas páginas 24, 25 e 65.

YATES, E.; YATES, L.; HARVEY, H. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. **Clinical Radiology**, v. 73, n. 9, p. 827–831, 2018. Citado 6 vezes nas páginas 21, 23, 40, 62, 63 e 64.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: [s.n.], 2014. (NIPS'14), p. 3320–3328. Citado na página 28.

ZECH, J. R.; BADGELEY, M. A.; LIU, M.; COSTA, A. B.; TITANO, J. J.; OERMANN, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. **PLOS Medicine**, Public Library of Science, v. 15, n. 11, p. 1–17, 11 2018. Disponível em: <https://doi.org/10.1371/journal.pmed.1002683>. Citado na página 44.

ZHANG, S.; AMAHONG, K.; SUN, X.; LIAN, X.; LIU, J.; SUN, H.; LOU, Y.; ZHU, F.; QIU, Y. The mirna: a small but powerful rna for covid-19. **Briefings in Bioinformatics**, v. 22, p. 1137–1149, 2021. Citado na página 17.

ZOPH, B.; VASUDEVAN, V.; SHLENS, J.; LE, Q. V. **Learning Transferable Architectures for Scalable Image Recognition**. 2018. Disponível em: <https://arxiv.org/abs/1707.07012>. Citado 4 vezes nas páginas 9, 30, 34 e 35.

ÇALLı, E.; SOGANCIOGLU, E.; van Ginneken, B.; van Leeuwen, K. G.; MURPHY, K. Deep learning for chest x-ray analysis: A survey. **Medical Image Analysis**, p. 102125, 2021. ISSN 1361-8415. Citado 6 vezes nas páginas 21, 24, 29, 30, 38 e 39.