



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

Justino Duarte Santos

**Diagnóstico de glomeruloesclerose usando comitês de
redes neurais convolucionais**

Orientador: Prof. Dr. Rodrigo de Melo Souza Veras

Teresina - PI
Outubro, 2024.

Justino Duarte Santos

**Diagnóstico de glomeruloesclerose usando comitês 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

Teresina - PI
Outubro, 2024.

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

S231d Santos, Justino Duarte.
Diagnóstico de glomeruloesclerose usando comitês de
redes neurais convolucionais / Justino Duarte Santos. --
2024.
88 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.”

1. Redes neurais. 2. Doença renal. 3. Diagnóstico
computadorizado. 4. Biópsia do rim. I. Veras, Rodrigo de
Melo Souza. II. Título.

CDD 006.3

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Justino Duarte Santos

Diagnóstico de glomeruloesclerose usando comitês 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. Romuere Rodrigues Veloso e Silva

Examinador Interno

Universidade Federal do Piauí

Prof. Dr. Vinicius Ponte Machado

Examinador Interno

Universidade Federal do Piauí

Prof. Dr. Pedro Pedrosa Rebouças Filho

Examinador Externo

Instituto Federal do Ceará

Prof. Dr. Lucas Ferrari de Oliveira

Examinador Externo

Universidade Federal do Paraná

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



Documento assinado digitalmente

RODRIGO DE MELO SOUZA VERAS

Data: 28/01/2025 16:35:46-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. Rodrigo de Melo Souza Veras

Orientador



Documento assinado digitalmente

IVAN SARAIVA SILVA

Data: 29/01/2025 08:34:07-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. Ivan Saraiva Silva

Coordenador

Teresina - PI, 07 de Outubro de 2024.

À minha família.

Agradecimentos

Em primeiro lugar, agradeço a Deus pelo dom da vida e da sabedoria.

Aos meus pais, irmãos e demais familiares, que sempre me apoiaram.

À minha querida esposa por todo o apoio, incentivo e amor.

Ao meu orientador prof. Dr. Rodrigo de Melo Souza Veras pela parceria, compreensão, ensinamentos e orientação.

A todos os amigos do LIMCI, e colaboradores do PathoSpotter.

À Universidade Federal do Piauí e Universidade Federal do Maranhão por oportunizar a qualificação profissional e desenvolvimento de pesquisa.

Ao Instituto Federal do Piauí por promover e incentivar a qualificação profissional dos docentes, possibilitando que se dediquem integralmente aos estudos de pós-graduação.

Agradeço a todos que de alguma forma contribuíram para minha formação acadêmica.

“O período de maior ganho em conhecimento e experiência é o período mais difícil da vida de alguém.”

Dalai Lama.

Resumo

A glomeruloesclerose é uma doença renal comum, caracterizada pela deposição de tecido cicatrizado que substitui o parênquima renal. Seu diagnóstico é baseado na análise histológica de biópsia da região cortical do rim por patologistas para indicar a presença e a extensão do dano. Para orientar o tratamento adequado e minimizar as chances da doença evoluir para estágios crônicos, o diagnóstico precoce é fundamental. No entanto, essa tarefa pode ser demorada e fatigante para o especialista. Nesse sentido os sistemas de diagnóstico auxiliado por computador servem ao médico oferecendo-lhe informações que podem ajudá-lo a diagnosticar doenças de forma mais rápida e menos trabalhosa. Para identificar glomérulo com esclerose, propomos um comitê de redes neurais convolucionais (*Convolutional Neural Networks* - CNN). Adaptamos cinco CNNs: DenseNet, EfficientNet, Inception, ResNet e VGG, com várias configurações de camadas totalmente conectadas. No total, foram avaliados 25 modelos diferentes, dos quais foram selecionados os membros votantes dos comitês formados. Avaliamos o treinamento com estágio único e com estratégia de dois estágios: primeiro treinamos somente a parte densa, depois foi feito um ajuste fino em toda a rede. Os experimentos foram realizados em três conjuntos de dados, compostos por 2.158 imagens ao todo, sendo 58% imagens de glomérulos diagnosticados com glomeruloesclerose e 42% saudáveis. Ainda aplicamos aumento de dados no conjunto de treinamento promovendo mais robustez ao modelo, resistência a *overfitting* e maior capacidade de generalização. Utilizando comitês de classificadores, nos conjuntos de testes, alcançamos acurácia média de 92,8%, AUC de 98,1% e kappa de 85,2%, considerado excelente. Entre as conclusões verificamos que o treinamento das CNNs em dois estágios é essencial para preservar transferência de aprendizado e o uso de comitês supera a performance das CNNs individualmente nessa tarefa de classificação.

Palavras-chave: Diagnóstico auxiliado por computador, Transferência de aprendizado, Doença renal, Análise de imagem de biópsia de rim, Comitê de CNNs.

Abstract

Glomerulosclerosis is a common kidney disease characterized by the deposition of scar tissue that replaces renal parenchyma. Its diagnosis is based on the histological analysis of biopsies from the renal cortex, performed by pathologists to assess the presence and extent of damage. Early diagnosis is crucial to guide appropriate treatment and minimize the risk of the disease progressing to chronic stages. However, this task can be time-consuming and exhausting for specialists. In this context, computer-aided diagnosis systems assist physicians by providing information that can help them diagnose diseases more quickly and with less effort. To identify glomeruli with sclerosis, we propose a committee of Convolutional Neural Networks (CNNs). We adapted five CNNs: DenseNet, EfficientNet, Inception, ResNet, and VGG, with various configurations of fully connected layers. In total, 25 different models were evaluated, from which the voting members of the formed committees were selected. We assessed both single-stage training and a two-stage strategy: first, we trained only the dense layers, followed by fine-tuning the entire network. The experiments were conducted on three data sources, comprising a total of 2,158 images, with 58% of the images diagnosed as glomeruli with glomerulosclerosis and 42% as healthy. We also applied data augmentation to the training set, enhancing the model's robustness, resistance to overfitting, and generalization capability. Using classifier committees, we achieved an average accuracy of 92.8%, AUC of 98.1%, and kappa of 85.2% on the test sets, which is considered excellent. Among our conclusions, we found that two-stage CNN training is essential to preserve transfer learning, and the use of committees outperforms the individual CNNs in this classification task.

Keywords: Computer-aided diagnosis, Transfer learning, Kidney disease, Kidney biopsy image analysis, CNN Ensemble.

Lista de ilustrações

Figura 1 – Anatomia do rim e do glomérulo.	21
Figura 2 – Segmentação de glomeruloesclerose	22
Figura 3 – Diagrama de conceitos e relação entre IA, ML e DL.	24
Figura 4 – Arquitetura genérica de uma CNN	26
Figura 5 – Operação de convolução.	26
Figura 6 – Operação de <i>max-pooling</i>	27
Figura 7 – Operação de <i>Global Average Polling</i>	27
Figura 8 – Arquitetura da CNN VGG-19.	29
Figura 9 – Esquema de bloco residual da CNN ResNet-50.	30
Figura 10 – Módulos Inception da Inception-V3.	31
Figura 11 – Esquema de bloco densamente conectado.	32
Figura 12 – Segmentos e camadas que formam a EfficientNet.	34
Figura 13 – Arquitetura da EfficientNet-B2.	35
Figura 14 – Curva ROC e AUC.	39
Figura 15 – Fluxograma dos experimentos.	45
Figura 16 – Amostra de imagens das bases utilizadas.	47
Figura 17 – Estratégia <i>k-fold</i> de validação cruzada para divisão do conjunto de dados.	48
Figura 18 – Exemplos das operações de aumento de dados.	50
Figura 19 – Modelagem de arquiteturas.	51
Figura 20 – Arquiteturas densas elaboradas.	51
Figura 21 – Estratégias de treinamentos aplicadas.	53
Figura 22 – Monitoramento da <i>loss</i> no conjunto de validação durante o processo de treinamento.	54
Figura 23 – Esquema de formação de comitês.	55
Figura 24 – Nomenclatura dos grupos formados nos comitês.	56
Figura 25 – Exemplo de aplicação das funções de apuração sem pesos (W_0).	59
Figura 26 – Exemplo de aplicação das funções de apuração com um peso por classificador (W_1).	61
Figura 27 – Exemplo de aplicação das funções de apuração com um peso para cada classe (W_2).	62
Figura 28 – Gráfico comparativo de AUC de validação nas etapas de treinamento.	70
Figura 29 – Métrica AUC obtidas pelos comitês no conjunto de validação.	71
Figura 30 – Comparação de métricas do <i>ensemble</i> 5Dd - P_2W_2 e de seus membros individuais no conjunto de testes.	73
Figura 31 – Amostra de imagens classificadas com o <i>ensemble</i> 5Dd - P_2W_2 no conjunto de testes.	74

Figura 32 – Mapas de atenção do GradCam das CNNs votantes no <i>ensemble</i> 5Dd - P_2W_2	74
Figura 33 – <i>Loss</i> e acurácia das CNNs do <i>ensemble</i> 5Dd - P_2W_2 em cada época.	76

Lista de tabelas

Tabela 1 – Arquitetura da Resnet-50.	30
Tabela 2 – Arquitetura da CNN Inception-V3.	32
Tabela 3 – Arquitetura da CNN DenseNet-201.	33
Tabela 4 – Matriz de confusão.	36
Tabela 5 – Níveis de concordância relativos às faixas de kappa.	37
Tabela 6 – Sumário dos trabalhos relacionados.	43
Tabela 7 – Quantidade de imagens na base de dados.	46
Tabela 8 – Quantidade de imagens nos subconjuntos de treino, validação e teste em uma execução do <i>k-fold</i>	48
Tabela 9 – Resumo das arquiteturas de CNNs avaliadas.	52
Tabela 10 – Parâmetros utilizados durante o treinamento das CNNs.	54
Tabela 11 – Resultados individuais com a estratégia OST.	64
Tabela 12 – Resultados das arquiteturas baseadas na CNN DenseNet-201.	66
Tabela 13 – Resultados das arquiteturas baseadas na CNN EfficientNet-B2.	67
Tabela 14 – Resultados das arquiteturas baseadas na CNN Inception-V3.	67
Tabela 15 – Resultados das arquiteturas baseadas na CNN ResNet-50.	68
Tabela 16 – Resultados das arquiteturas baseadas na CNN VGG-19.	68
Tabela 17 – Aumento médio das métricas com o treinamento da etapa DFT em relação às obtidas na SFT.	69
Tabela 18 – <i>Ensembles</i> formados e seus membros classificadores.	71
Tabela 19 – Resultados dos comitês do grupo 5Dd com as funções de apuração tipo P_2 no conjunto de testes.	72
Tabela 20 – Comparação do método proposto com outros trabalhos relacionados.	75

Lista de abreviaturas e siglas

acc	Acurácia
AUC	Área sob a curva ROC (<i>Area Under the Curve</i>)
BN	Normalização em lote (<i>Batch Normalization</i>)
CAD	Diagnóstico Auxiliado por Computador (<i>Computer-Aided Diagnosis</i>)
CNN	Rede Neural Convolucional (<i>Convolutional Neural Network</i>)
DFT	<i>Deep Fine-tuning</i>
DL	Aprendizagem Profunda (<i>Deep Learning</i>)
DLM	Doença de Lesões Mínimas
dsn	DenseNet-201
efn	EfficientNet-B2
espec	Especificidade
FN	Falso Negativo
FP	Falso Positivo
FPR	Taxa de Falsos Positivos (<i>False Positive Rate</i>)
GAP	<i>Global Average Pooling</i>
GESF	Glomeruloesclerose Segmentar e Focal
GRAD-CAM	<i>Gradient-weighted Class Activation Map</i>
GS	Glomeruloesclerose
H&E	Hematoxilina-Eosina
IA	Inteligência Artificial
icp	Inception-V3
ILSVRC	<i>ImageNet Large Scale Visual Recognition Challenge</i>
KNN	<i>K-Nearest Neighbour</i>

ML	Aprendizagem de máquina (<i>Machine Learning</i>)
MLP	<i>Multi Layer Perceptron</i>
ND	Nefropatia Diabética
NPS	Normal ou Parcialmente Esclerosado
npv	Valor Preditivo Negativo (<i>Negative Predictive Value</i>)
OST	<i>One-step tuning</i>
ppv	Valor Preditivo Positivo (<i>Positive Predictive Value</i>)
PAS	Ácido Periódico de Schiff (<i>Periodic Acid-schif</i>)
PMJ	<i>Pratas Metenamina e de Jones</i>
PGL	Lesão Proliferativa Glomerular (<i>Prolifetive Glomerular Lesion</i>)
ReLu	<i>Rectified Linear Unit</i>
RF	Floresta Aleatória (<i>Random Forest</i>)
RNA	Rede Neural Artificial
ROC	<i>Receiver Operating Characteristic</i>
rsn	Resnet-50
sensib	sensibilidade
SFT	<i>Shallow Fine-tuning</i>
SVM	Máquina de Vetor de Suporte (<i>Suport Vector Machine</i>)
TL	Transferência de aprendizado (<i>Transfer learning</i>)
vgg	VGG-19, (<i>Visual Geometry Group Network</i>)
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XAI	Inteligência Artificial Explicável (<i>Explainable Artificial Intelligence</i>)

Sumário

1	INTRODUÇÃO	16
1.1	Objetivo Geral	17
1.2	Objetivos Específicos	18
1.3	Contribuições e Produções Científicas	18
1.4	Organização do Trabalho	19
2	FUNDAMENTAÇÃO TEÓRICA	20
2.1	O rim e o glomérulo	20
2.2	Biópsia renal	21
2.3	Glomeruloesclerose	22
2.4	Etapas do processamento digital de imagens.	22
2.5	Inteligência artificial, aprendizagem de máquina e <i>deep learning</i>	23
2.5.1	Classificadores baseados em comitês (<i>Ensembles</i>)	25
2.5.2	Redes neurais convolucionais	25
2.5.3	VGG	29
2.5.4	ResNet	29
2.5.5	Inception	31
2.5.6	DenseNet	31
2.5.7	EfficientNet	33
2.5.8	Explicabilidade de modelos <i>deep learning</i>	34
2.6	Métricas de avaliação	36
2.7	Considerações Finais	39
3	TRABALHOS RELACIONADOS	40
3.1	Considerações Finais	44
4	MATERIAIS E MÉTODOS	45
4.1	Conjuntos de dados (dataset)	46
4.1.1	Validação Cruzada	47
4.1.2	Aumento de dados e pré-processamento	48
4.2	Modelagem de CNNs	50
4.3	Estratégias de treinamento	53
4.4	Formação de Comitês	54
4.5	Considerações Finais	62
5	RESULTADOS E DISCUSSÃO	64

5.1	Resultados com CNNs individuais	64
5.1.1	Resultados com treinamento em etapa única (OST)	64
5.1.2	Resultados com treinamento em duas etapas (SFT e DFT)	65
5.2	Seleção de modelos e formação de <i>ensembles</i>	69
5.3	Resultados com <i>ensembles</i>	71
5.4	Considerações Finais	77
6	CONCLUSÃO	78
6.1	Trabalhos Futuros	79
	REFERÊNCIAS	81

1 Introdução

Os rins são um par de órgãos localizados na parte posterior do abdômen e são protegidos pela caixa torácica. Esses órgãos são cercados por uma fina camada de tecido conjuntivo e por uma camada de gordura. Além disso, são responsáveis por filtrar o sangue das artérias renais, retirando o excesso de água, sais e resíduos, excretando-os pela urina. São também órgãos que desempenham um papel fundamental no metabolismo da vitamina D, na fabricação de hormônios que estimulam a produção de hemácias e na regulação da pressão arterial ([SOCIETY, 2017](#); [FUND, 2019](#)).

A glomeruloesclerose (GS) é uma doença caracterizada por alterações morfológicas de minúsculos vasos sanguíneos denominados glomérulos - presentes nos rins ([RISDON; TURNER, 2012](#); [BUENO et al., 2020](#)). Nos glomérulos, resíduos e fluidos são filtrados do sangue e removidos do corpo através da urina. Tanto crianças como adultos podem ser diagnosticados com GS, e estima-se que quase 20% dos adultos com doença renal desenvolvam a glomeruloesclerose ([DAVITA INC., 2023b](#)).

Doenças glomerulares ou glomerulonefrites são a terceira maior causa de doenças renais em estágio terminal no Brasil ([SESSO et al., 2016](#)). O censo brasileiro de diálise 2020 aponta que as glomerulonefrites foram causa de 9% a 10% dos casos de pacientes em diálise ([NERBASS et al., 2022](#)) entre 2017 e 2020. Um levantamento global aponta que a glomeruloesclerose figura entre as glomerulonefrites predominantes em quatro continentes, na América do Norte alcançam quase 40% dos casos ([O'SHAUGHNESSY et al., 2017](#)).

Assim como em outras doenças, quanto maior a demora no diagnóstico da GS, menores são as chances de o paciente preservar a função renal. Para diagnosticar doenças renais com precisão, é necessário realizar a análise histológica de uma biópsia renal. No entanto, fazer um diagnóstico preciso via biópsia geralmente é demorado, mesmo para patologistas treinados, observando-se, em alguns casos, discordância entre especialistas quanto ao diagnóstico.

Diante disso, percebe-se que o processamento computadorizado das imagens de biópsia para apoiar essa tarefa médica, traz melhorias potenciais na eficiência desses procedimentos em patologia renal, contribuindo também para um diagnóstico mais objetivo e padronizado, principalmente em hospitais ou países com poucos nefro-patologistas.

Avanços na aquisição e processamento de imagens digitais tornaram possível usar computacionalmente os dados fornecidos a partir de imagens médicas. Isso levou a um considerável interesse no desenvolvimento de sistemas automáticos de diagnóstico

médico para melhorar os serviços prestados pela comunidade médica. Esses sistemas são chamados de sistemas de diagnóstico auxiliado por computador (*Computer-Aided Diagnosis* - CAD) (DOI, 2007).

Estudos sobre sistemas CAD mostram que eles podem auxiliar os médicos a medir estruturas anatômicas importantes (KOTYK et al., 2016), acompanhar estágios de evolução de doenças (GINLEY et al., 2018), diagnosticar e planejar o melhor tratamento (MARSH et al., 2018; SILVA et al., 2023). Esses sistemas podem também aliviar o trabalho repetitivo dos profissionais, prevenindo erros decorrentes de fadiga e facilitando tarefas como a análise de imagens, a contagem de elementos biológicos, ou a localização de regiões de interesse.

De acordo com Yi et al. (2021), a integração de inteligência artificial e medicina clínica permite o desenvolvimento de ferramentas e recursos para apoiar a tomada de decisão clínica. Vários aspectos da nefrologia serão afetados por essas novas tecnologias. A patologia renal é um bom exemplo onde se pode esperar melhorias nos fluxos de trabalho e análises histológicas, embora ainda haja muito trabalho a ser feito para validar esses novos métodos (HUO et al., 2021).

Algoritmos de aprendizagem de máquina, especialmente aqueles baseados em Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN), têm alcançado excelente desempenho na classificação e análise de imagens e são aplicados em diversas áreas de pesquisa, inclusive na medicina (YU et al., 2021). Essas técnicas adquiriram atenção considerável em histologia e patologia (LAAK; LITJENS; CIOMPI, 2021).

Considerando a importância do diagnóstico da glomeruloesclerose e o potencial dos sistemas CAD, neste trabalho desenvolvemos um método baseado em aprendizagem profunda (*Deep Learning* - DL) para classificar imagens de biópsia renal em saudáveis ou com glomeruloesclerose.

1.1 Objetivo Geral

O objetivo desse trabalho é desenvolver um sistema CAD para auxiliar o diagnóstico de glomeruloesclerose. O sistema recebe como entrada imagens histológicas de glomerulos e as classifica como saudáveis ou com GS. O método deve ser robusto frente a presença de ruídos e a variações decorrentes do processo de aquisição das imagens, também deve ser abrangente quanto aos corantes utilizados e necessitar da mínima intervenção do usuário ou configuração de parâmetros.

1.2 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, faz-se necessário atingir os seguintes objetivos específicos:

- Desenvolver modelos *deep learning* baseados em redes neurais convolucionais, adequados para lidar com imagens de biópsia renal no diagnóstico de glomeruloesclerose.
- Aplicar aumento de dados sobre as imagens de treinamento para dar mais robustez aos modelos e resistência a ruídos.
- Avaliar diferentes abordagens de treinamento dos modelos, testá-los e analisar o desempenho separadamente.
- Definir critérios para seleção dos melhores modelos para formar classificadores baseados em comitês com vistas a obter melhor performance em relação aos modelos individuais.
- Analisar a performance dos comitês formados, comparando-os entre si, com seus membros e com trabalhos semelhantes da literatura.

1.3 Contribuições e Produções Científicas

As principais contribuições do presente trabalho são enumeradas a seguir:

1. Desenvolvimento de um método baseado em comitês de CNNs para classificar imagens de glomérulos saudáveis ou com esclerose;
2. Verificação do impacto do treinamento de CNNs em dois estágios e em estágio único sobre a transferência de aprendizado;
3. Avaliação de 25 arquiteturas de CNNs individuais;
4. Formação e avaliação de comitês, utilizando critérios de seleção de membros votantes e mecanismos de apuração da classificação final.

A pesquisa de doutorado gerou publicações, incluindo artigos em periódicos, apresentações em congressos e capítulos de livros. Na lista a seguir destacamos as publicações diretamente relacionadas a este trabalho:

1. **SANTOS, J. D.**; MACHADO, V. P.; OLIVEIRA, L. R.; SANTOS, W. L. C.; ALDEMAN, N. L.; DUARTE, A. A.; VERAS, R. M. S.. Glomerulosclerosis Identification Using a

- Modified Dense Convolutional Network. In: **Brazilian Conference on Intelligent Systems**. Cham: Springer International Publishing, 2022. p. 237-252.
2. **SANTOS, J. D.**; SILVA, R. R. V. ; OLIVEIRA, L. R.; SANTOS, W. L. C.; ALDEMAN, N. L.; DUARTE, A. A.; VERAS, R. M. S.. Glomerulosclerosis detection with pre-trained CNNs ensemble. **Computational Statistics**, v. 39, n. 2, p. 561-581, 2024.
 3. CLARO, M. L.; VOGADO, L. H. S.; **SANTOS, J. D.**; VERAS, R. M. S.. Utilização de técnicas de data augmentation em imagens: Teoria e prática. Sociedade Brasileira de Computação, 2020.
 4. SILVA, L; MACHADO, V. P.; VERAS, R. M. S.; AITA, K.; do MONTE, S.; ALDEMAN, N. L.; **SANTOS, J. D.**. Using Clinical Data and Deep Features in Renal Pathologies Classification. In: **International Conference on Intelligent Systems Design and Applications**. Cham: Springer Nature Switzerland, 2023. p. 137-146.
 5. ROCHA, Ana A. F.; BORGES, Rodrigo N.; BATISTA, Rodrigo E. C.; NASCIMENTO Rhaylson S.; MORICONY, Emery F.; **SANTOS, Justino D.**; VERAS, Rodrigo M. S.. Seleção de Backbone Para Extração de Características com a U-Net na Segmentação de Patologias Renais. in Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI). SBC 2024.

1.4 Organização do Trabalho

Os demais capítulos deste trabalho foram organizados da seguinte forma:

- O Capítulo 2 trata dos conceitos fundamentais necessários para a construção desta pesquisa.
- O Capítulo 3 apresenta um resumo de trabalhos relacionados à análise computacional de imagens de histologia renal.
- O Capítulo 4 descreve as etapas dos experimentos realizados para se chegar ao método proposto.
- O Capítulo 5 apresenta os resultados obtidos com métodos avaliados nos experimentos.
- Finalmente, o Capítulo 6 apresenta as considerações finais e sugestões de trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo abordaremos conceitos fundamentais necessários para compreensão do tema abordado no presente trabalho. Vamos detalhar aspectos da seara médica: a estrutura anatômica do rim e do glomérulo (Seção 2.1), o exame de biópsia renal (Seção 2.2) e a glomeruloesclerose (Seção 2.3). E posteriormente alguns aspectos da área da computação utilizados nesta pesquisa: processamento digital de imagens (Seção 2.4), inteligência artificial, aprendizagem de máquina e *deep learning* (Seção 2.5) e métricas de avaliação (Seção 2.6). Por fim são apresentadas as considerações finais na Seção 2.7.

2.1 O rim e o glomérulo

Os rins são órgãos bilaterais em forma de feijão, de cor marrom-avermelhada e localizados na região posterior do abdômen (Figura 1-a). Geralmente, o rim direito é ligeiramente menor e mais baixo que o esquerdo devido à presença do fígado. Cada rim pesa cerca de 125-170 gramas nos homens e 115-155 gramas nas mulheres, e são revestidos por uma cápsula renal fibrosa e resistente. Além disso, duas camadas de gordura servem como proteção (NEWMAN, 2019).

Os rins são os principais responsáveis pela manutenção do equilíbrio químico de nossos corpos, nossa sobrevivência depende do funcionamento normal destes órgãos vitais. Na região cortical dos rins estão presentes milhões de microestruturas que são consideradas suas unidades funcionais — os **néfrons** (Figura 1-b). Neles se realiza o processo de filtração do sangue que permite que este se livre das impurezas ou resíduos não aproveitáveis pelo organismo e os elimine através da urina. No néfron se encontra o **glomérulo** - um conjunto de pequenos vasos sanguíneos emaranhados que se encontram envoltos pela cápsula de Bowman (Figura 1-c). É no glomérulo que chega o sangue para ser filtrado (FERRON; RANCANO, 2007) e portanto ele constitui elemento fundamental para a manutenção das funções renais.

As glomerulopatias são doenças que acometem os glomérulos. Elas possuem diferentes subtipos histopatológicos. No Brasil, as glomerulopatias encontram-se entre as principais causas de doença renal terminal, representando 11% dos pacientes em diálise (COSTA et al., 2017). Em muitas situações, o nefrologista precisa lançar mão da biópsia renal para determinar o tipo de glomerulopatia com precisão (Sociedade Brasileira de Nefrologia, 2023).

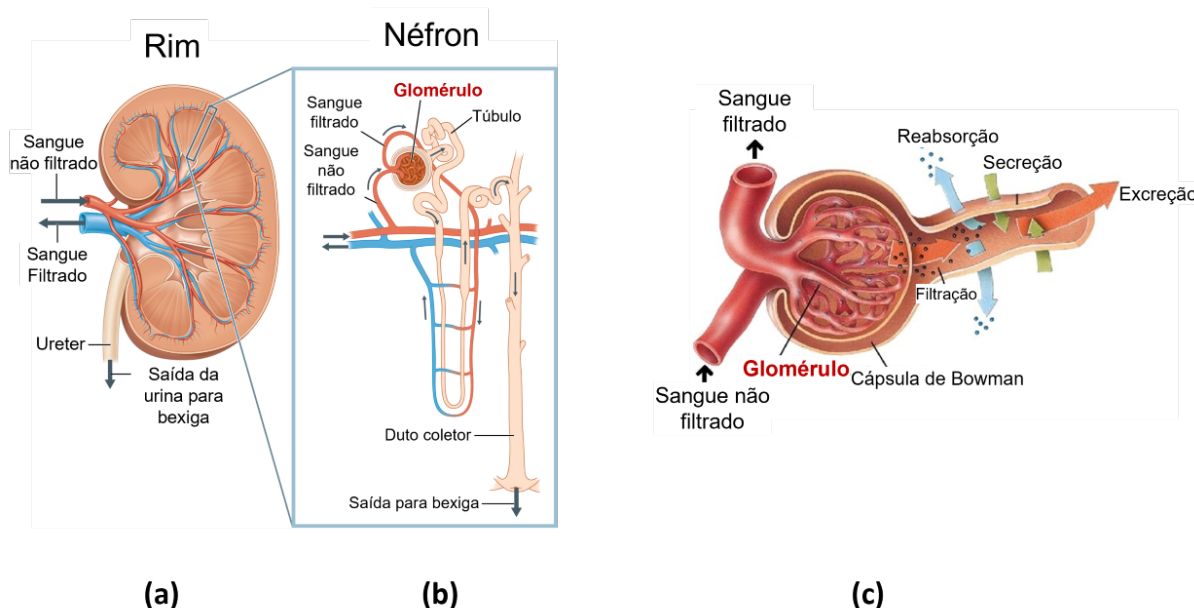


Figura 1 – Ilustração da anatomia do rim e do glomérulo. Em (a) ilustramos o rim em corte, em (b) temos em destaque o néfron e em (c) uma visualização destacada do glomérulo. Fonte: adaptado de <www.niddk.nih.gov/health-information/kidney-disease/nephrotic-syndrome-adults> e de <www.anatomiaemfoco.com.br/sistema-urinario/rins>

2.2 Biópsia renal

Uma biópsia renal é um teste em que pequenos pedaços de tecido renal são coletados e examinados ao microscópio. O procedimento geralmente é feito no hospital com anestesia local ou geral. Uma biópsia renal geralmente leva cerca de uma hora, durante a qual o médico localizará os rins, limpará a área, administrará anestesia e coletará amostras de tecido (DAVITA INC., 2023a).

O material biológico coletado no exame é processado em laboratório, passa por etapas de fixação, coloração com corantes químicos específicos, e é cortado com auxílio de um micrótomo em fatias muito finas que são acomodadas em lâminas para análise no microscópio.

A biópsia renal é muitas vezes necessária para diagnosticar doenças renais com precisão. Além de estabelecer o diagnóstico, a biópsia permite avaliar a cronicidade e a progressão para definir o prognóstico e escolher a terapia adequada (DHAUN et al., 2014). Além disso, um critério para aceitar ou rejeitar rins transplantados depende muito da determinação da análise realizada pelo patologista de fatores como as porcentagens de glomérulos normais e com esclerose (MARSH et al., 2018).

2.3 Glomeruloesclerose

A glomeruloesclerose (GS) é uma glomerulopatia na qual se observam alterações morfológicas, como cicatrização ou endurecimento em regiões dos glomérulos (BUENO et al., 2020) prejudicando seu correto funcionamento. A identificação precoce dessa doença é de fundamental importância para preservação das funções renais do paciente.

O diagnóstico da GS é obtido por meio da análise de biópsia renal. A Figura 2 apresenta imagens histológicas de glomérulos onde foram destacadas regiões com glomeruloesclerose. Observa-se principalmente alterações na textura da região, mas há formas variadas de manifestação.

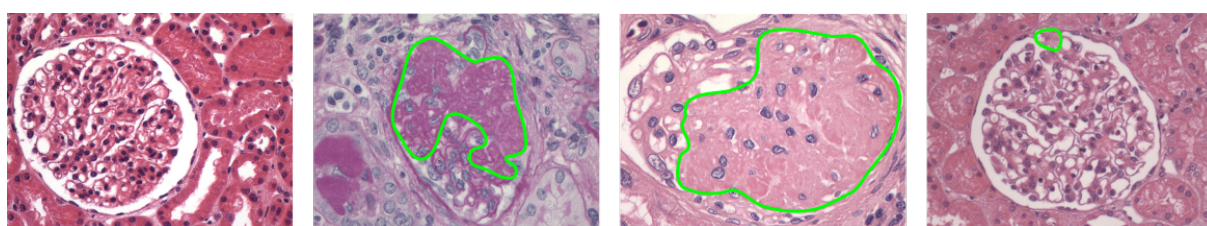


Figura 2 – Segmentação de glomeruloesclerose. A imagem da esquerda é de um glomérulo saudável. Nas demais, a área delimitada pela linha verde indica as regiões do glomérulo acometidas pela GS.

Nos estágios iniciais da glomeruloesclerose, uma pessoa pode não apresentar nenhum sintoma, semelhante à doença renal em estágio inicial. Quando alguém tem glomeruloesclerose, a cicatrização pode alterar o processo de filtração nos rins, permitindo que as proteínas passem do sangue para a urina, o que é chamado de proteinúria. A proteinúria ocorre quando grandes quantidades de proteína são perdidas na urina e é identificada por meio de exames laboratoriais. Os sinais de proteinúria com grande perda de proteína incluem: tornozelos e mãos inchados; inchaço dos olhos; retenção de líquido abdominal; urina espumosa (DAVITA INC., 2023b).

2.4 Etapas do processamento digital de imagens.

O processamento digital de imagens é definido como um conjunto de técnicas computacionais que transformam uma imagem de entrada em uma saída desejada (GONZALEZ; WOODS, 2008). Esse conjunto de técnicas permite extrair e identificar informações de imagens e melhorar a qualidade visual de aspectos estruturais, facilitando a percepção humana e a interpretação automática por meio de máquinas (PEDRINI; SCHWARTZ, 2008).

Existem vários algoritmos com propósitos muito específicos, que juntos formam a metodologia final. Portanto, uma metodologia difere de outra na forma como compõe suas etapas ou ferramentas, mas normalmente segue as etapas apresentadas por Gonzalez e

Woods (2008):

1. Aquisição das imagens: as imagens são capturadas por meio de um dispositivo ou sensor e transformadas em um formato que pode ser processado posteriormente. Porém, como resultado desse processo, a imagem resultante pode apresentar falhas devido às condições de iluminação ou às características do dispositivo, por exemplo;
2. Pré-processamento: o objetivo é melhorar a qualidade da imagem por meio de técnicas para reduzir o ruído, realçar o contraste e suavizar certas estruturas da imagem;
3. Segmentação: consiste em localizar e extrair áreas de interesse contidas na imagem, com o objetivo de dividir a imagem isolando os diferentes objetos que a compõem;
4. Representação e descrição: denominada de extração de características, pois extrai informações que podem ser usadas para discriminar classes de objetos;
5. Reconhecimento e interpretação: reconhecimento é o processo que atribui um identificador aos objetos da imagem. A interpretação consiste em atribuir um significado ao conjunto de objetos reconhecidos.

2.5 Inteligência artificial, aprendizagem de máquina e *deep learning*

Os termos que intitulam esta seção atualmente estão bem difundidos, apesar de não carregarem estritamente o mesmo significado, são muitas vezes aplicados de forma indiscriminada a depender do contexto. Resumimos no diagrama da Figura 3 os seus respectivos conceitos e a relação entre eles.

A Inteligência Artificial (IA) é um termo abrangente que contempla dentro de si qualquer técnica computacional que leve a máquina a mimetizar o comportamento inteligente do homem. Apesar de estar em evidência nos dias atuais nas mais variadas e surpreendentes aplicações, o termo em si data de 1950. Evidentemente que muitos avanços neste campo têm acontecido, potencializados também graças ao aumento da capacidade de hardware.

Algoritmos tradicionais programam os computadores de forma explícita, ou seja, o desenvolvedor insere através de uma linguagem de programação o passo a passo exato que o computador deve seguir para resolver diretamente um determinado problema.

Outro formato se observa quanto ao uso de técnicas de IA de aprendizagem de máquina (*machine learning* - ML). Neste caso o computador passa a aprender as regras,

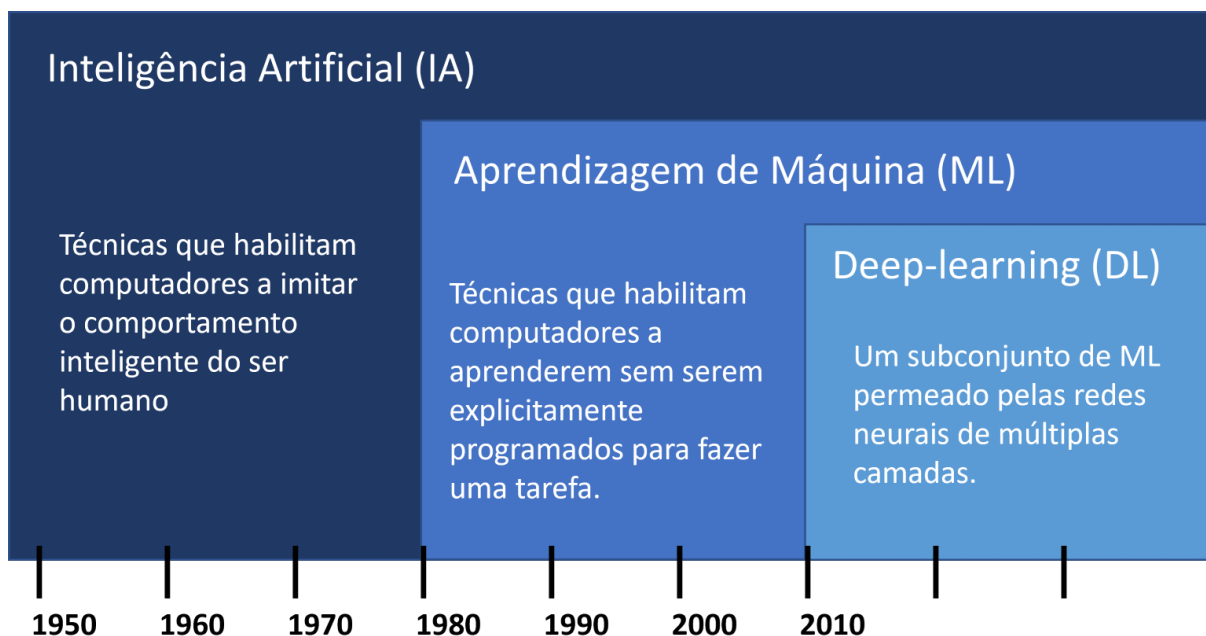


Figura 3 – Diagrama de conceitos e relação entre IA, ML e DL.

reconhecer os padrões ou aproximar a função desejada. O programador se encarrega não mais de inserir as regras diretamente, mas de cuidar do algoritmo de aprendizagem.

Um subconjunto de ML são os algoritmos de *deep learning*. Seu arcabouço está centrado em redes de unidades inspiradas no funcionamento dos neurônios do cérebro. A sessão 2.5.2 abordará de forma mais abrangente sobre essa tecnologia.

O tipo de aprendizagem que tratamos nessa proposta é aprendizagem supervisionada: o algoritmo observa alguns exemplos de pares de entrada e saída, e aprende uma função que faz o mapeamento da entrada para a saída (RUSSELL; NORVIG, 2016).

Algoritmos supervisionados de aprendizagem de máquina utilizam modelos estatísticos para fazer previsões. Esses modelos são construídos (na etapa de treino) com base em exemplos cuja variável a ser predita já é conhecida. É importante ressaltar que tradicionalmente o algoritmo é planejado para realizar sua tarefa para domínios similares nos conjuntos de treino e teste. Uma mudança de domínio, a princípio, necessitaria de uma reconstrução do modelo (novo treinamento). Em contraste, utilizar transferência de aprendizado permite que os domínios, tarefas e distribuições usados no treinamento e teste sejam diferentes (PAN; YANG, 2010; RIBANI; MARENGONI, 2019).

Um problema que pode ocorrer com os modelos de ML é o *overfitting*. Acontece quando um modelo aprende os detalhes e ruídos nos dados de treinamento. Isso significa que o ruído ou flutuações aleatórias nos dados de treinamento são captados e aprendidos como conceitos pelo modelo. A questão é que esses conceitos não se aplicam a novos dados e impactam negativamente a capacidade de generalização dos modelos. Em outras palavras ocorre um superajuste do modelo aos dados de treinamento.

2.5.1 Classificadores baseados em comitês (*Ensembles*)

O comitê ou *ensemble* de classificadores é uma técnica que precede bastante o surgimento do paradigma de aprendizagem profunda (DASARATHY; SHEELA, 1979). A teoria na qual está baseada é bastante simples e baseia-se na noção de *sabedoria das multidões*: em vez de confiar em apenas um modelo de previsão, um conjunto de modelos (pré-treinados) é criado e as previsões desses modelos são combinadas em uma classificação final (CLARO, 2022).

A ideia original foi desenvolvida para reduzir a variância dos classificadores para obter um melhor desempenho geral, conforme explica Dietterich (2000), os comitês geralmente apresentam performance melhorada em relação aos classificadores individuais. Podemos encontrar na literatura outros trabalhos que usaram *ensembles* com sucesso em sistemas CAD (ALVES; FILHO; KRITSKI, 2019; PICCIALLI et al., 2021; CLARO, 2022).

De maneira geral, a produção de um *ensemble* envolve três tarefas principais: a produção dos classificadores de base, a seleção de membros do comitê e a definição do mecanismo de combinação das previsões (ZHOU, 2012). Na seleção de membros, observa-se que a variedade dos classificadores escolhidos direciona para maior robustez e capacidade de generalização do modelo, pois cada membro classifica com base em aspectos e modos diferentes. A finalidade dos mecanismos de combinação, ou funções de apuração, é receber a saída de cada membro de um comitê e combiná-las, produzindo, assim, uma única saída de classificação consolidada.

2.5.2 Redes neurais convolucionais

Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN) (LECUN et al., 1998) são comumente aplicadas no campo de aprendizagem de máquina para classificação de imagens. Uma grande vantagem dessa técnica é que elas são capazes de detectar automaticamente as características importantes, e permitem a criação de uma arquitetura de tamanho viável para o processamento do grande volume de dados (pixels) que imagens digitais contêm.

A arquitetura de uma CNN tipicamente inclui duas seções. A primeira é formada por uma sequência de operações de convolução seguidas de operações de *pooling*, a segunda parte é composta de camadas totalmente conectadas similares a redes neurais convencionais do tipo *multilayer perceptron* (MLP) (MCCULLOCH; PITTS, 1943). A Figura 4 ilustra a arquitetura genérica de uma CNN. A rigor, existem outras topologias com fluxos residuais e camadas de normalização, por exemplo. Por conveniência, chamaremos a parte inicial da rede até o ponto onde os mapas são linearizados como parte convolucional.

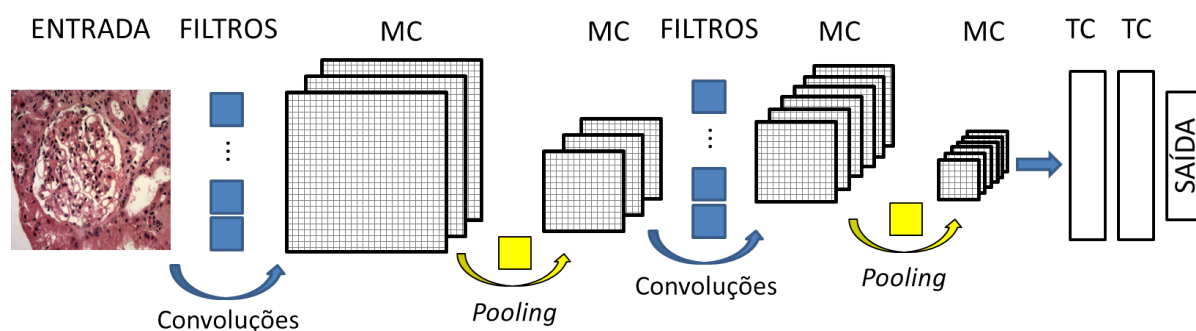


Figura 4 – Arquitetura genérica de uma CNN. Os quadrados azuis e amarelos significam filtros de convolução e *pooling* respectivamente. MC = mapa de características, TC = camada totalmente conectada.

Uma operação de convolução sobre uma imagem consiste em efetuar uma varredura sobre a mesma, visitando cada um dos seus pixels e multiplicando sua vizinhança por uma matriz de pesos (filtro), o somatório desses produtos para cada pixel tomado gera um mapa de características de dimensões semelhantes à da imagem de origem. Em cada camada convolucional de uma CNN muitos filtros diferentes podem ser aplicados, cada um deles gera um mapa diferente. A Figura 5 ilustra esta operação.

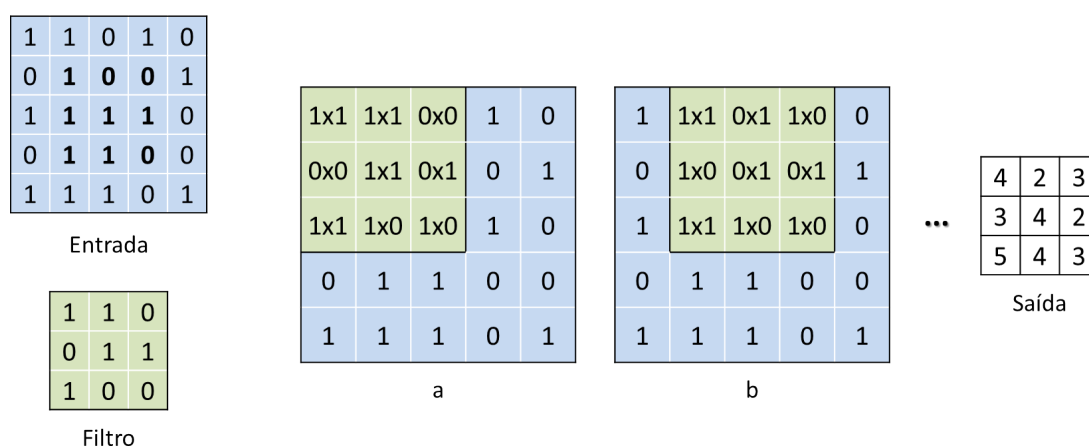


Figura 5 – Operação de convolução. O filtro com os pesos “visita” os pixels da entrada (neste exemplo ilustramos apenas para os 9 pixels centrais - em negrito), em a e b ilustramos o filtro centrados em pixels diferentes, em cada posição os pesos são multiplicados, o somatório desses produtos compõem cada um dos valores da saída. $\sum a = 4$, $\sum b = 2$, como se vê nos dois primeiros valores do mapa de saída.

A operação de *Pooling* segue um mecanismo parecido ao da convolução, mas em vez de usar filtros de pesos, a janela de *pooling* sintetiza as informações por meio de alguma função, por exemplo o filtro *max-pooling*, produz como resultado o maior valor de cada vizinhança. Geralmente, na operação de *pooling* a janela não é centrada em cada um dos pontos da matriz de entrada, pois se desloca sobre ela com passos maiores (*stride* > 1), isso faz com que os mapas gerados na saída tenham dimensões

menores, consequentemente também diminui o número de parâmetros da rede e o tempo computacional para treiná-la. A Figura 6 exemplifica o funcionamento da operação de *max-pooling*.

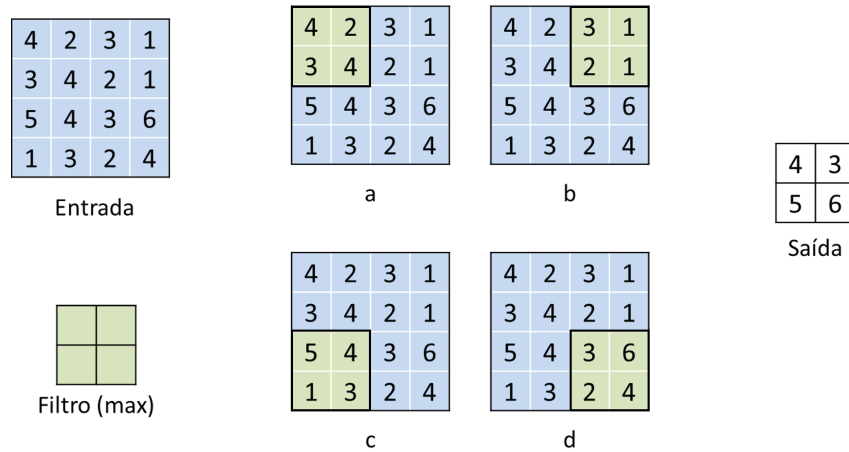


Figura 6 – Operação de *max-pooling*. O valor máximo é tomado em cada vizinhança. Em a, b, c e d, ilustramos o filtro em diferentes posições. Neste exemplo o filtro anda com “passos” de duas casas ($stride=2 \times 2$), produzindo um mapa de saída de dimensões menores.

Outro mecanismo de *pooling*, o global, este geralmente aplicado antes da parte densa, também reduz a dimensionalidade dos mapas de características. No entanto, a operação é aplicada em todo o mapa de uma vez e gera um único valor para representá-lo. A Figura 7 exemplifica o funcionamento do *Global Average Pooling* (GAP), que realiza uma operação de média sobre os valores de cada mapa.

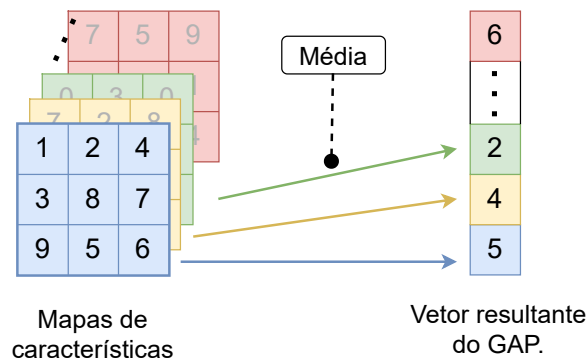


Figura 7 – Operação de *Global Average Polling*. O valor médio de cada mapa é calculado e um vetor unidimensional é produzido contendo esses valores.

Outro elemento essencial em redes neurais são as funções de ativação, que desempenham um papel crucial na introdução de não-linearidades aos modelos. Sem elas, uma rede neural seria apenas uma combinação linear de suas entradas, incapaz de capturar e modelar a complexidade de dados reais. As funções de ativação, como ReLU (Rectified Linear Unit), sigmoid e *softmax*, transformam as saídas das camadas

anteriores e permitem a propagação de características complexas através das camadas subsequentes.

A ReLU (Equação 2.1) é popular por sua simplicidade e eficiência computacional, definindo valores negativos como zero e mantendo os positivos inalterados. Enquanto a sigmoid (Equação 2.2) é uma função suave que comprime os valores de saída no intervalo de 0 a 1. Já a *softmax* (Equação 2.3) é uma função de ativação usada especialmente na camada de saída de modelos de classificação. Ela transforma os logits (saídas das camadas anteriores) em probabilidades, garantindo que a soma das probabilidades seja igual a 1.

$$ReLU(x) = \max(0, x). \quad (2.1)$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}. \quad (2.2)$$

$$softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}, \quad (2.3)$$

onde x_i é o i -ésimo elemento do vetor de entrada, e a soma no denominador é feita sobre todos os elementos do vetor.

Após a sequência de convoluções e *pooling*, geralmente os mapas de características gerados são dados como entrada para camadas totalmente conectadas. Nesta parte, a arquitetura e mecanismo de operação é similar às redes neurais tradicionais, e na última camada tem-se os dados de saída (classificação).

As CNNs têm sido amplamente utilizadas em aprendizado de máquina, especialmente em imagens médicas. Com suas arquiteturas profundas de múltiplas camadas, as CNNs podem mapear características de imagem em diferentes níveis de abstração e têm sido usadas com sucesso no desenvolvimento de ferramentas de diagnóstico médico (KALANTARI et al., 2018; OLIVEIRA et al., 2022), muitas vezes superando em acurácia os métodos convencionais de extração de atributos (TAJBAKHSH et al., 2016).

No contexto de avaliação de desempenho de CNNs, a *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) (RUSSAKOVSKY et al., 2015) – uma competição anual para avaliar a capacidade de modelos de aprendizado de máquina em reconhecer e classificar objetos em imagens de grande escala – foi um marco na pesquisa de redes neurais convolucionais e promoveu avanços significativos na precisão e na eficiência dos algoritmos de reconhecimento de imagem. O dataset de classificação ImageNet consiste em aproximadamente 1,3 milhão de imagens de treino, 50 mil imagens de validação e 100 mil imagens de teste, distribuídas em 1.000 classes distintas. A seguir descreveremos

algumas arquiteturas de CNNs já estabelecidas na literatura e que estão disponíveis publicamente já com os seus pesos treinados na base de dados ImageNet.

2.5.3 VGG

As arquiteturas das CNNs VGG, propostas por [Simonyan e Zisserman \(2015\)](#), são do tipo sequencial e recebem imagens RGB de 224×224 pixels como entrada. Originalmente, foram desenvolvidas seis arquiteturas contendo de 11 a 19 camadas treináveis, organizadas em 5 blocos convolucionais e um bloco final com camadas totalmente conectadas. Entre cada bloco uma operação de *max-pooling* reduz a dimensionalidade dos mapas de características. A Figura 8 apresenta a arquitetura da CNN VGG-19, nela os filtros de convolução são todos de dimensão 3×3 e o filtro de *max-pooling* tem dimensão e *stride* 2×2 , que reduz a largura e altura dos mapas de características pela metade.

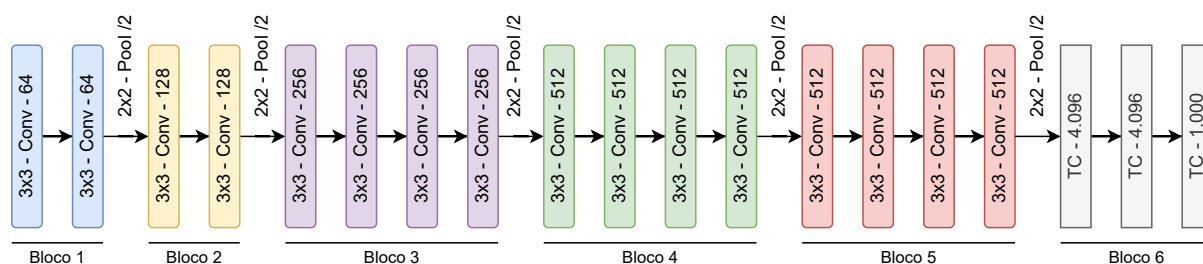


Figura 8 – Arquitetura da CNN VGG-19. Em cada camada, anotamos dados sobre a dimensão e quantidade de filtros. Ex. 3×3 - Conv - 64 indica uma camada convolucional de 64 filtros de tamanho 3×3 . TC - 1.000 equivale a mil neurônios totalmente conectados com a camada adjacente.

2.5.4 ResNet

O efeito de gradiente desvanecente em redes neurais convolucionais (CNNs) ocorre quando os gradientes utilizados para atualizar os pesos da rede durante o treinamento tornam-se muito pequenos, isso acontece especialmente em redes profundas. Devido aos gradientes pequenos, as atualizações dos pesos são insignificantes, tornando o treinamento lento e, em casos extremos, fazendo com que a rede pare de aprender. Esse efeito dificulta a aprendizagem eficaz das camadas mais profundas da rede, prejudicando a sua capacidade de capturar padrões complexos nos dados.

Uma das inovações principais da ResNet ([HE et al., 2015](#)) é o uso de conexões residuais, essa técnica consiste basicamente em transportar o mapa de características de um ponto da rede para um ponto posterior, por meio de um fluxo paralelo que não passa por todas as camadas convolucionais intermediárias e assim preserva os dados do mapa. As conexões residuais ajudam a mitigar o problema do gradiente desvanecente, permitindo a construção de redes mais profundas sem degradação no desempenho.

Na ResNet-50, a arquitetura está organizada em blocos residuais, cada bloco usa uma estrutura que inclui convoluções 1x1 que reduzem a dimensionalidade (profundidade) do mapa de características, seguidas por convoluções 3x3 e, finalmente, convoluções 1x1, que possuem maior quantidade de filtros e assim restauram a dimensionalidade anterior (Figura 9). Na conexão residual ocorre a soma do mapa de características presente no início do bloco com o produzido na sua saída. A arquitetura da ResNet-50 é composta por 50 camadas, 1 camada convolucional inicial e 48 camadas convolucionais organizadas em 16 blocos residuais, e uma camada densa na saída conforme apresentado na Tabela 1.

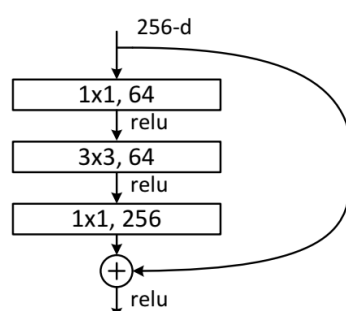


Figura 9 – Esquema de bloco residual da CNN ResNet-50. Fonte: [He et al. \(2015\)](#)

Tabela 1 – Arquitetura da Resnet-50.

Nome	Dimensão	Leiaute da arquitetura	OBS
Conv1	112×112	7×7 (64) , <i>stride</i> =2	Dimensão de entrada: 224×224 .
Conv2	56×56	$\begin{bmatrix} 1 \times 1 (64) \\ 3 \times 3 (64) \\ 1 \times 1 (256) \end{bmatrix} \times 3$	Previamente, em Conv2 a dimensão é reduzida via <i>max-pooling</i> 3×3 com <i>stride</i> =2.
Conv3	28×28	$\begin{bmatrix} 1 \times 1 (128) \\ 3 \times 3 (128) \\ 1 \times 1 (512) \end{bmatrix} \times 4$	Em Conv3, Conv4 e Conv5, a dimensão é reduzida na primeira camada convolucional do primeiro bloco residual aplicando convolução com <i>stride</i> =2.
Conv4	14×14	$\begin{bmatrix} 1 \times 1 (256) \\ 3 \times 3 (256) \\ 1 \times 1 (1.024) \end{bmatrix} \times 6$	
Conv5	7×7	$\begin{bmatrix} 1 \times 1 (512) \\ 3 \times 3 (512) \\ 1 \times 1 (2.048) \end{bmatrix} \times 3$	
Saída	1×1	GAP, TC-1000	<i>Global Average Pooling</i> seguido de camada totalmente conectada com mil neurônios.

OBS: Em cada camada é descrito o tamanho dos filtros e a quantidade deles entre parênteses. Os blocos residuais são exibidos entre colchetes (v. Figura 9) e ao lado há a quantidade de blocos empilhados na arquitetura.

2.5.5 Inception

A Inception ([SZEGEDY et al., 2015](#)) é uma arquitetura de rede neural convolucional desenvolvida pela Google, projetada para performar bem mesmo sob restrições de memória e recursos computacionais. Essa arquitetura é caracterizada pela utilização de módulos Inception (Figura 10), que permitem a aplicação de convoluções com diferentes tamanhos em paralelo. Esses módulos ajudam a capturar informações de vários tamanhos de filtro, mantendo o número de parâmetros relativamente baixo.

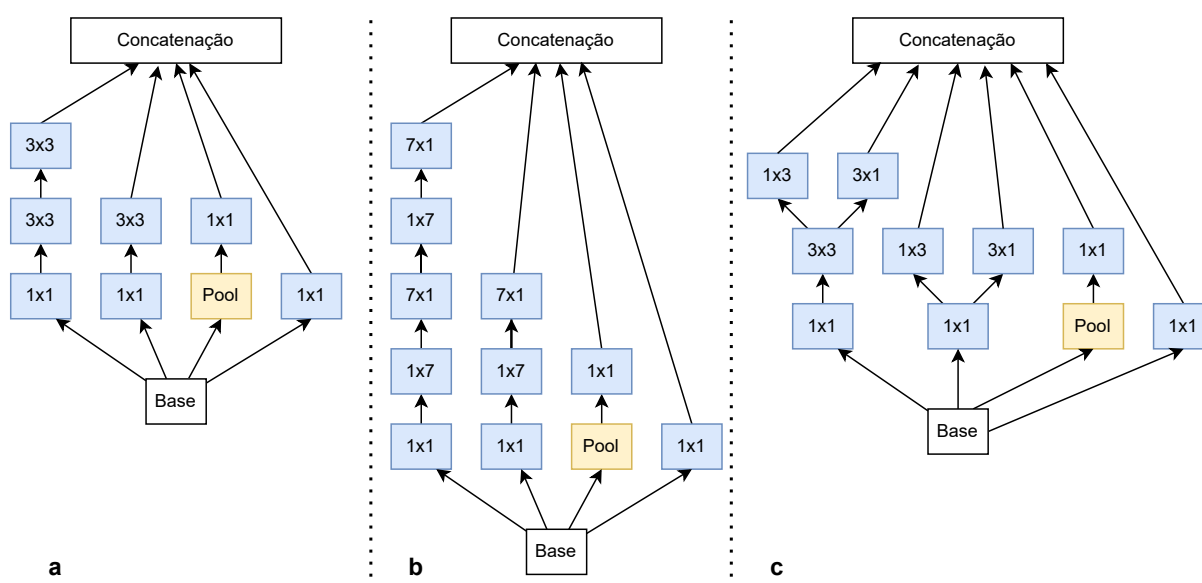


Figura 10 – Módulos Inception da CNN Inception-V3 ([SZEGEDY et al., 2016](#)). Os retângulos azuis representam filtros de convolução e seus respectivos tamanhos.

A arquitetura Inception contribuiu significativamente para o avanço em tarefas de reconhecimento de imagens e é amplamente utilizada em várias aplicações de visão computacional. Entre os destaques dessas arquiteturas, a versão Inception-V3 ([SZEGEDY et al., 2016](#)) implementa convoluções fatoradas e normalização em lote (*Batch Normalization* - BN), o que aumentou ainda mais a eficiência computacional e a precisão. A Tabela 2 detalha a estrutura da Inception-V3.

2.5.6 DenseNet

A DenseNet (*Densely Connected Convolutional Network*) é uma arquitetura de rede neural convolucional proposta por [Huang, Liu e Weinberger \(2016\)](#) que atacou o problema do efeito de gradiente desvanecente. Outras abordagens, como a da ResNet, criam fluxos alternativos mais curtos das camadas iniciais, para algumas camadas posteriores. A DenseNet utiliza um padrão de conectividade simples, que visa garantir o máximo fluxo de informações entre as camadas da rede.

Tabela 2 – Arquitetura da CNN Inception-V3.

Tipo	Tamanho do filtro / <i>stride</i>	Dimensão de entrada
Convolução	3 x 3 / 2	299 x 299 x 3
Convolução	3 x 3 / 1	149 x 149 x 32
Conv. <i>padded</i>	3 x 3 / 1	147 x 147 x 32
Pool	3 x 3 / 2	147 x 147 x 64
Convolução	3 x 3 / 1	73 x 73 x 64
Convolução	3 x 3 / 2	71 x 71 x 80
Convolução	3 x 3 / 1	35 x 35 x 192
3 x Inception	Conforme Figura 10-a	35 x 35 x 288
3 x Inception	Conforme Figura 10-b	17 x 17 x 768
3 x Inception	Conforme Figura 10-c	8 x 8 x 1.280
Pool	8 x 8	8 x 8 x 2.048
Linear	logits	1 x 1 x 2.048
Softmax	Classificação/saída	1 x 1 x 1.000

OBS: A dimensão de saída de cada camada é a mesma da entrada da camada seguinte. Conv. *padded* adiciona um contorno à entrada, assim o filtro pode ser centralizado nos pixels das bordas. Entre os blocos Inception, a redução de dimensionalidade é obtida aplicando *stride*=2x2 nas camadas que antecedem a concatenação (v. Figura 10).

Dessa forma, a principal inovação da DenseNet é a conexão densa, onde cada camada recebe a saída de todas as camadas anteriores como entrada, observando a correspondência dos tamanhos dos mapas de características. Sua arquitetura é organizada em blocos densos (Figura 11), agrupando várias camadas que concatenam suas saídas para serem passadas às camadas subsequentes. A arquitetura da versão DenseNet-201 é descrita na Tabela 3.

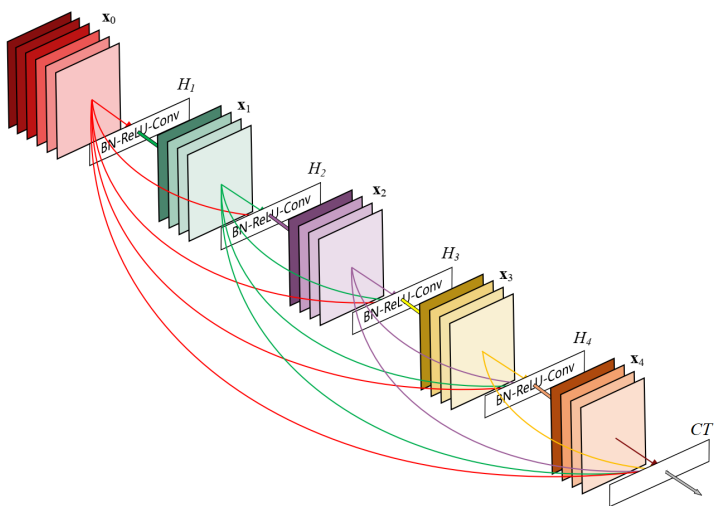


Figura 11 – Esquema de bloco densamente conectado. X_i são mapas de características e H_i contém as operações de convolução. CT indica uma camada de transição que conecta a saída ao bloco seguinte. Fonte: Huang, Liu e Weinberger (2016), adaptado.

Tabela 3 – Arquitetura da CNN DenseNet-201.

Nome	Dimensão de saída	Leiaute da arquitetura
Convolução	$112 \times 112 \times 64$	7×7 (64) conv, <i>stride</i> =2, BN + ReLU
Pooling	$56 \times 56 \times 64$	3×3 max-pool, <i>stride</i> =2
Bloco densamente conectado 1	$56 \times 56 \times 256$	$\begin{bmatrix} 1 \times 1 \text{ (128) BRConv} \\ 3 \times 3 \text{ (32) BRConv} \\ \text{Concatenação} \end{bmatrix} \times 6$
Transição 1	$56 \times 56 \times 128$	1×1 (128) BRConv
	$28 \times 28 \times 128$	2×2 avg-pool, <i>stride</i> =2
Bloco densamente conectado 2	$28 \times 28 \times 512$	$\begin{bmatrix} 1 \times 1 \text{ (128) BRConv} \\ 3 \times 3 \text{ (32) BRConv} \\ \text{Concatenação} \end{bmatrix} \times 12$
Transição 2	$28 \times 28 \times 256$	1×1 (256) BRConv
	$14 \times 14 \times 256$	2×2 avg-pool, <i>stride</i> =2
Bloco densamente conectado 3	$14 \times 14 \times 1.792$	$\begin{bmatrix} 1 \times 1 \text{ (128) BRConv} \\ 3 \times 3 \text{ (32) BRConv} \\ \text{Concatenação} \end{bmatrix} \times 48$
Transição 3	$14 \times 14 \times 896$	1×1 (896) BRConv
	$7 \times 7 \times 896$	2×2 avg-pool, <i>stride</i> =2
Bloco densamente conectado 4	$7 \times 7 \times 1.920$	$\begin{bmatrix} 1 \times 1 \text{ (128) BRConv} \\ 3 \times 3 \text{ (32) BRConv} \\ \text{Concatenação} \end{bmatrix} \times 32$
Classificação	$1 \times 1 \times 1.920$	7×7 GAP
	1.000	TC-1.000, <i>softmax</i>

OBS: A entrada tem tamanho $224 \times 224 \times 3$; (*n*) **BRConv**: sequência de BN + ReLU + *n* filtros de convolução; **avg-pool**: pooling via média; **TC-1.000**: Camada não convolucional totalmente conectada de mil neurônios. Cada operação de **concatenação** empilha os 32 mapas resultantes da convolução 3×3 aos que foram a entrada da BRConv 1×1 anterior.

2.5.7 EfficientNet

Tendo observado que a ampliação de CNNs é largamente utilizada para obter melhor acurácia dos modelos, os pesquisadores [Tan e Le \(2019\)](#) observaram que a maneira mais comumente aplicada para este fim, promovia o aumento da arquitetura em profundidade (quantidade de camadas), largura (quantidade de canais) ou resolução de imagem (dimensões dos mapas). O cerne do seu trabalho foi então estudar e repensar o processo de ampliação de CNNs, buscando uma maneira não arbitrária de fazê-la, com base em princípios que reduzissem o trabalho de definições e ajustes manuais.

Nesse trabalho foi proposta a EfficientNet, uma família de modelos de redes neurais convolucionais, cujas arquiteturas iniciam em uma arquitetura base a partir da qual são produzidos outros modelos maiores utilizando o princípio da escalabilidade composta, que equilibra de forma otimizada a profundidade, largura e resolução da rede. Em vez de simplesmente aumentar um desses fatores sem critérios definidos, na EfficientNet aplica-se um método composto para redimensionar esses três fatores simultaneamente.

As arquiteturas EfficientNet são montadas a partir de estruturas sequenciais, aqui denominadas segmentos, que servem de base para formação de estruturas mais complexas. Esses segmentos estão dispostos na Figura 12. Nos segmentos 1 e 3 há uma camada denominada *Depthwise convolution* (DW - Conv), essa é uma convolução que aplica um filtro convolucional separado a cada canal de entrada, ao contrário da convolução normal (Conv), que mescla os canais para gerar a saída. Outra camada que destacamos é a *dropout*, ela desativa aleatoriamente uma fração dos neurônios durante o treinamento, o que força a rede a aprender redundâncias e a se tornar mais robusta e resistente a *overfitting*.

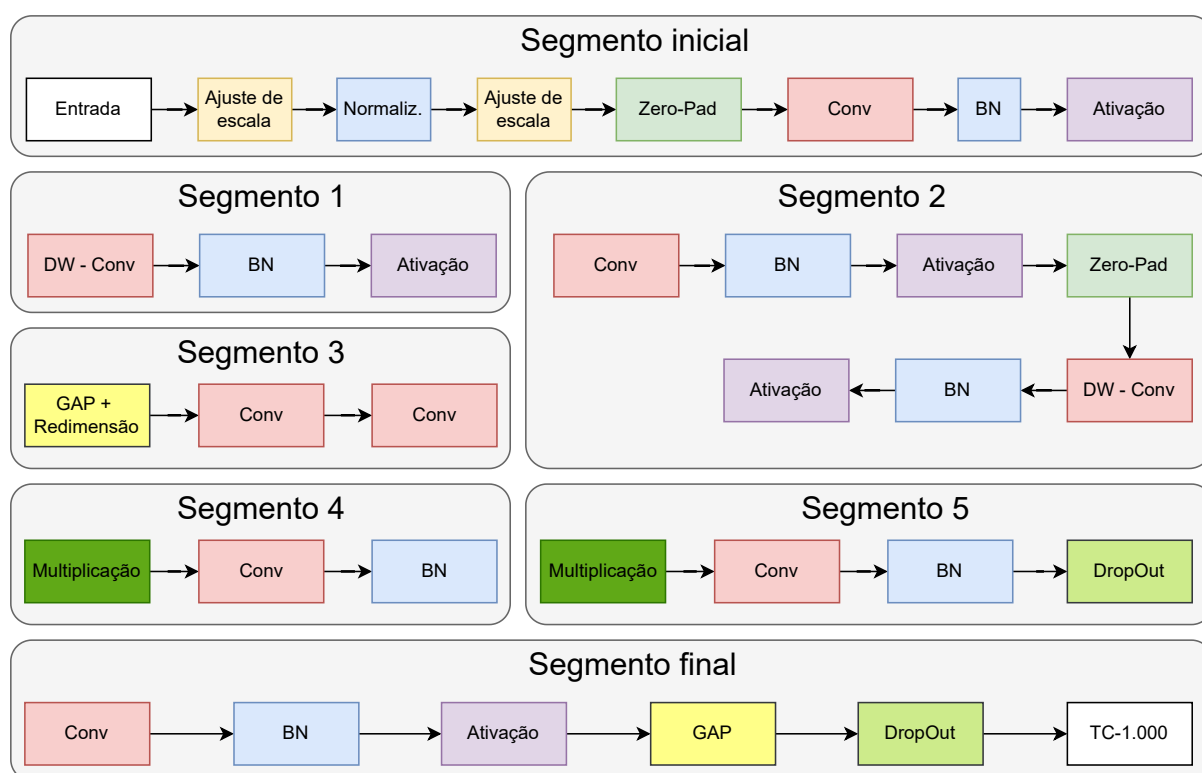


Figura 12 – Segmentos e camadas que formam a EfficientNet.

Os segmentos são dispostos sob fluxos não sequenciais e formam assim módulos com os quais se monta a arquitetura final da EfficientNet. Esta por sua vez está organizada em 7 blocos localizados entre os segmentos inicial e final. Na Figura 13, ilustramos a arquitetura da versão EfficientNet-B2, e os módulos formados a partir dos segmentos já mencionados.

2.5.8 Explicabilidade de modelos *deep learning*

Os algoritmos de aprendizagem de máquina baseados em *deep learning*, logicamente se incluem aqui as CNNs, padecem da característica de serem intrinsecamente do tipo “caixa-preta” pois não conseguimos saber como todos os neurônios individuais trabalham juntos para chegar ao resultado final. Em outras palavras, uma rede neural é

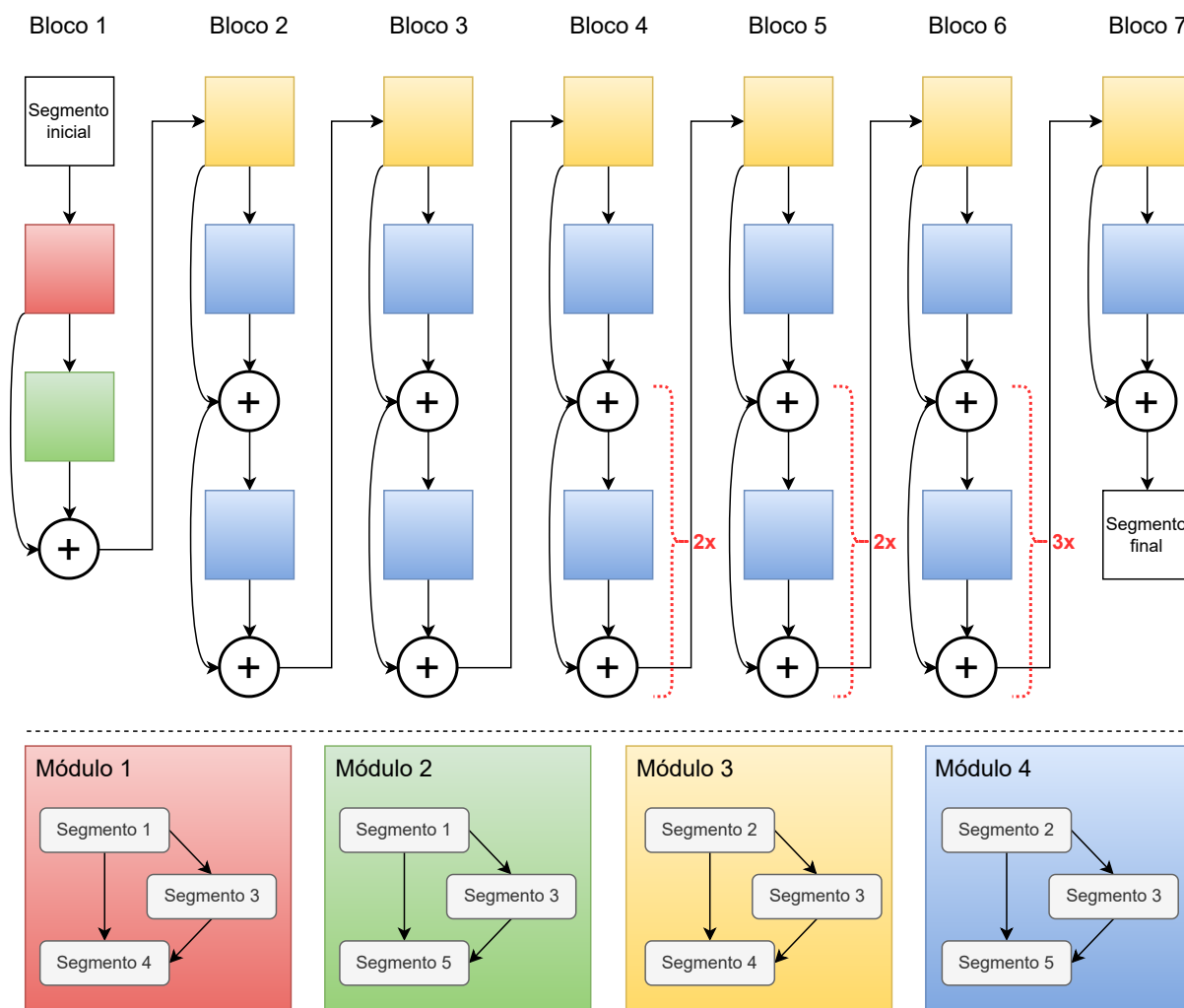


Figura 13 – Arquitetura da EfficientNet-B2. A rede inicia e finaliza com os segmentos inicial e final (v. Figura 12). Entre eles, a arquitetura é organizada em 7 blocos, cada um deles é estruturado em módulos conforme detalhamento na parte inferior da figura. O bloco 1 é composto pelos módulos 1 e 2. Nos demais blocos, o módulo 3 é o inicial e o módulo 4 atua como elemento de ligação para os demais módulos.

uma caixa-preta no sentido de que, embora possa aproximar uma função qualquer, estudar sua estrutura não fornecerá nenhuma percepção sobre a estrutura da função que está sendo aproximada.

O cerne da questão está na compreensão humana do aprendizado adquirido pelo modelo. Geralmente são arquiteturas densas, que contêm milhões de parâmetros, e múltiplas camadas e interligações. E a cada nível de aprofundamento da rede se torna mais difícil associar um significado humanamente compreensível aos mapas de características gerados pelas convoluções, por exemplo.

Tais nuances fazem parte do campo de pesquisa de explicabilidade de IAs (*Explainable Artificial Intelligence* - XAI). O desenvolvimento de métodos e estudos que permitam explicar as decisões de uma rede neural profunda tornou-se, assim, uma área

ativa e ampla de pesquisa (RAS; GERVEN; HASELAGER, 2018; RAS et al., 2022; MONTAVON; SAMEK; MÜLLER, 2018; ZHANG; ZHU, 2018).

A depender do contexto de aplicação, pode ser de menor importância compreender o funcionamento do método (recomendação de conteúdo online, por exemplo), mas em outros cenários (como o diagnóstico de doenças) é importante que além de boas taxas de acerto, se tenha algum nível dessa compreensão. Neste sentido, Ras et al. (2022) em sua pesquisa introduz a seguinte questão: “Como uma pessoa que é responsabilizada por uma decisão pode confiar na recomendação de uma rede neural profunda e justificar seu uso?” – A confiança e a justificação dificilmente serão alcançadas se o usuário não tiver acesso a uma explicação satisfatória para o processo que levou à recomendação.

Entre as iniciativas que apresentam métodos de explicabilidade, há técnicas que se aplicam bem ao caso de imagens e CNN, é o caso do *Gradient-weighted Class Activation Map* (GRAD-CAM) (SELVARAJU et al., 2017). Esta técnica de XAI consiste em gerar mapas de atenção sobre a imagem, baseados nos gradientes da rede. Através deles pode-se visualizar regiões específicas da imagem que apresentaram maior contribuição para a classificação de uma determinada classe.

2.6 Métricas de avaliação

A matriz de confusão confronta os resultados previstos e reais para o mesmo conjunto de testes. Existem quatro valores nessa matriz: o verdadeiro positivo (VP), que indica o número de imagens corretamente classificadas como GS; o verdadeiro negativo (VN), correspondendo ao número de classificações corretas de imagens saudáveis; o falso positivo (FP), representando o número de imagens classificadas como GS, mas que são saudáveis; e, por fim, o falso negativo (FN), que se refere ao número de imagens erroneamente classificadas como saudáveis.

A Tabela 4 ilustra esses quatro quadrantes de classificações possíveis. A partir da matriz de confusão, é possível calcular diversas métricas capazes de indicar o desempenho dos classificadores.

Tabela 4 – Matriz de confusão.

Classificação Computada	Classificação Real	
	Saudável	Glomeruloesclerose
Saudável	VN	FN
Glomeruloesclerose	FP	VP

A acurácia (acc) é uma métrica bastante conhecida e simples, ela mede a taxa de acerto geral na classificação para ambas as classes de maneira unificada; ela é obtida pela razão entre o número de imagens classificadas corretamente e o total, de acordo com a Equação 2.4.

$$acc = \frac{VP + VN}{VP + VN + FP + FN}. \quad (2.4)$$

A acurácia, embora seja uma métrica bem difundida e intuitiva para avaliar o desempenho de um modelo de classificação, apresenta algumas limitações, especialmente em conjuntos de dados desbalanceados. O coeficiente kappa (Eq. 2.5) (COHEN, 1960), por sua vez, oferece uma medida mais robusta da concordância entre as predições do modelo e os valores reais, pois leva em consideração a probabilidade esperada (P_e) dos avaliadores concordarem com a classificação, com base na distribuição dos dados.

$$kappa = \frac{A - P_e}{1 - P_e}, \quad (2.5)$$

onde,

$$P_e = P_{GS} + P_{Saudavel}, \quad (2.6)$$

$$P_{GS} = \frac{VP + FN}{N} * \frac{VP + FP}{N}, \quad (2.7)$$

$$P_{Saudavel} = \frac{VN + FN}{N} * \frac{VN + FP}{N}, \quad (2.8)$$

e,

$$N = VP + VN + FP + FN. \quad (2.9)$$

O valor máximo do kappa é 100%, o que indica a concordância perfeita entre os avaliadores. Para outros valores, a Tabela 5, proposta por Landis e Koch (1977), é comumente utilizada como forma de manter uma nomenclatura consistente para descrever grau de concordância associada aos valores de kappa.

Tabela 5 – Níveis de concordância relativos às faixas de kappa.

kappa (%)	Nível de concordância
<0	Sem acordo
0 – 20	Insignificante
20 – 40	Baixo
40 – 60	Moderado
60 – 80	Significativo
80 – 100	Quase perfeito

A acurácia e o kappa são calculados considerando todos os valores da matriz de confusão, mas para algumas análises é interessante observar taxas separadamente a depender da classe predita. Nesse conjunto se enquadram as métricas *positive predictive value* (ppv) também conhecida como precisão - taxa de acerto entre as instâncias classificadas como positivas; e a *negative predictive value* (npv) - taxa de acerto entre as

instâncias classificadas como negativas; calculadas conforme as Equações 2.10 e 2.11, respectivamente.

$$ppv = \frac{VP}{VP + FP}, \quad (2.10)$$

$$npv = \frac{VN}{VN + FN}, \quad (2.11)$$

De maneira análoga às métricas ppv e npv, o par de métricas sensibilidade ou *recall* (Equação 2.12) e especificidade (Equação 2.13) indicam as taxas de acerto a depender da classe, porém o denominador aqui considera a classificação real e não a predita. Dessa maneira, a sensibilidade (*sensib*) indica a taxa de acerto entre as instâncias que são positivas e a especificidade (*espec*) representa a taxa de acerto entre instâncias que são negativas.

$$sensib = \frac{VP}{VP + FN}. \quad (2.12)$$

$$espec = \frac{VN}{VN + FP}. \quad (2.13)$$

Sensibilidade e especificidade são duas métricas fundamentais na avaliação de testes diagnósticos, especialmente na área médica. Mas para uma avaliação mais objetiva entre modelos de classificação geralmente é desejável ter uma métrica unificada. Ao encontro disso, temos a AUC (*Area Under the Curve*), ela representa a capacidade do modelo de distinguir entre classes positivas e negativas.

A AUC é uma métrica mais elaborada e não é calculada com base apenas nos valores da matriz de confusão. Para tanto, a AUC é obtida a partir da Curva ROC (*Receiver Operating Characteristic*). Essa curva é um gráfico que mostra a sensibilidade em relação à taxa de falsos positivos ($FPR = 1 - espec$) para diferentes limiares de classificação.

A saída de um modelo de classificação (predição), seja binária ou categórica, é frequentemente representada por um valor numérico ou um vetor que indica uma probabilidade para cada classe. A partir de um determinado limiar definido, a classe de saída é decidida matematicamente, verificando se a probabilidade estimada (predita) atinge o limite estabelecido pelo limiar. A curva ROC é construída aplicando limiares diferentes e para cada um deles se obtém o par FPR e sensibilidade. A AUC é a área sob essa curva (Figura 14).

Quanto maior a AUC, melhor é o classificador. Quando seu valor é 100%, significa que o modelo apresentou um desempenho perfeito, classificando corretamente todos os

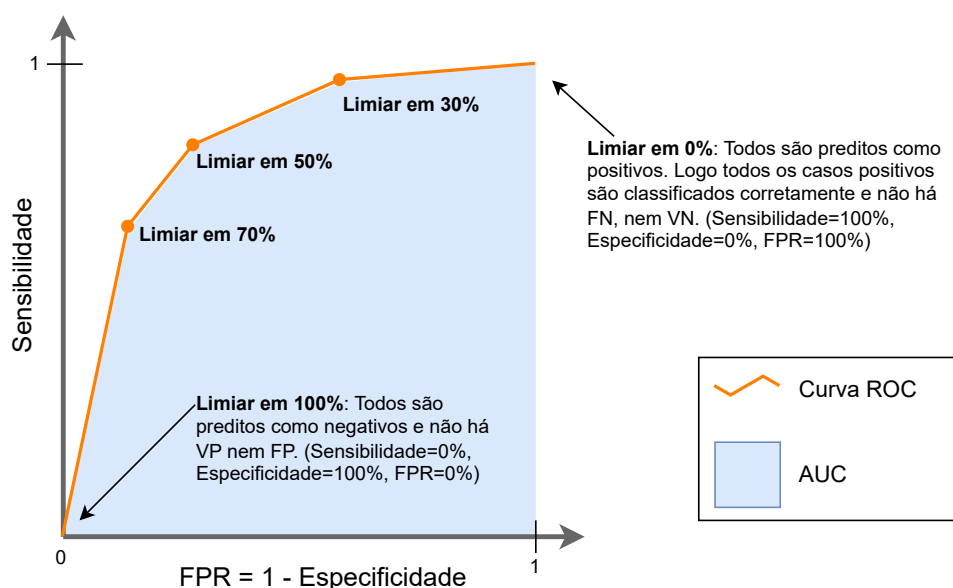


Figura 14 – Curva ROC e AUC. O classificador gera como saída a probabilidade de a entrada ser da classe positiva. Um limiar define a partir de que valor a saída deverá ser considerada positiva. A curva ROC é produzida a partir da avaliação de diferentes limiares aplicados ao classificador; para cada limiar, a FPR e a sensibilidade são calculadas e adiciona-se o ponto correspondente no gráfico. A sequência desses pontos formam a curva ROC. A área sob essa a curva é a métrica AUC.

exemplos. Quando é de 50% significa que o modelo tem um desempenho aleatório, não consegue distinguir entre as classes.

2.7 Considerações Finais

Neste capítulo, foi apresentada a fundamentação teórica necessária para a compreensão do presente trabalho. Abordamos tanto aspectos da área médica quanto da computação, além disso dispomos sobre as métricas utilizadas para descrição e avaliação de desempenho dos modelos analisados.

3 Trabalhos Relacionados

Na literatura, existem estudos que tratam de temas próximos ao abordado nesta proposta de tese, no âmbito de métodos computacionais para auxiliar o diagnóstico de doenças renais através da análise de imagens histológicas.

Muitos trabalhos aplicam processamento de imagem e técnicas de inteligência artificial no campo da patologia renal ([SANTOS et al., 2022](#)). Alguns deles analisam túbulos, vasos sanguíneos, e interstício ([BEVILACQUA et al., 2017](#); [KOLACHALAMA et al., 2018](#); [ZHENG et al., 2021a](#)). No entanto, grande parte dos estudos focam no glomérulo, pois geralmente é nessa estrutura onde ocorrem muitos achados histológicos essenciais para o diagnóstico de diversas doenças.

Para realizar o diagnóstico computadorizado de glomerulopatias, a partir da digitalização da imagem de lâmina inteira com amostras de tecido renal (biópsia), o primeiro passo no procedimento é a detecção (localização/segmentação) de glomérulos. Este problema também já é objeto de estudos recentes. Algumas soluções propostas usam métodos que empregam várias características definidas manualmente ([SHEEHAN; KORSTANJE, 2018](#)) ou fazem uso de CNNs ([GINLEY et al., 2019](#); [HECKENAUER et al., 2020](#); [BUENO et al., 2020](#); [SHUBHAM et al., 2021](#); [REHEM et al., 2021](#); [SILVA et al., 2022](#); [SAMANT et al., 2023](#)). Discorremos a seguir sobre relevantes estudos presentes no estado da arte, que classificam os achados patológicos nas imagens glomerulares.

[Barros et al. \(2017\)](#) propuseram um sistema computacional para detectar lesões glomerulares proliferativas (*Prolifetive Glomerular Lesion* - PGL), diferenciando-as de imagens saudáveis. O conjunto de dados utilizado era composto de 811 imagens de glomérulos individuais, tratados com os corantes Hematoxilina-Eosina (H&E) ou Periodic Acid-schif (PAS). O método utilizado pelos autores inclui uma etapa de pré-processamento de imagem e segmentação. Como descritores utilizaram a contagem de regiões com núcleos e a aglomeração deles. O algoritmo *k-nearest neighbor* (KNN) ([AHA; KIBLER; ALBERT, 1991](#)) foi utilizado como classificador. A acurácia alcançada no conjunto de validação foi de 85%.

No estudo apresentado por [Ginley et al. \(2018\)](#), os autores propuseram uma abordagem para definir a progressão estrutural da nefropatia diabética (ND) em glomérulos humanos. Os autores segmentaram os compartimentos glomerulares e quantificaram 47 características de cada glomérulo, incluindo informações de textura, morfologia, medidas de distâncias e características anotadas manualmente. O conjunto de dados consistiu em 514 imagens de glomérulos obtidas de 13 biópsias humanas diagnosticadas com ND e cinco tecidos renais humanos sem anormalidades histológicas.

Eles treinaram modelos baseados em Naive Bayes ([HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#)) para realizar a classificação. Os estágios patológicos da ND vão de I onde as lesões são mínimas, seguidos pelos estágios IIa, IIb, III e IV, no qual a doença já está avançada ([TERVAERT et al., 2010](#)). Nos resultados eles relataram que seu método distinguiu o estágio patológico IIa do estágio III com sensibilidade/especificidade 0,89/0,93, e o estágio IIb do estágio III com sensibilidade/especificidade 0,7/0,8.

Assim como [Barros et al. \(2017\)](#), e usando a mesma base de dados, [Chagas et al. \(2020\)](#) trabalharam na detecção de PGL. No entanto, além da classificação binária (lesionada ou saudável), sua proposta também realizava a classificação em subcategorias específicas de PGL: endocapilar, mesangial e ambas. Esses autores construíram uma arquitetura baseada em CNN para extrair características de imagens de glomérulo, e esses atributos foram fornecidos a um classificador SVM. O método proposto alcançou uma precisão de 82%.

[Zheng et al. \(2021b\)](#) propuseram um sistema baseado em CNN que detecta e classifica achados patológicos glomerulares na nefrite lúpica. Eles usaram um conjunto de dados de 349 imagens de lâmina inteira de biópsia tratadas com PAS nas quais localizaram 3.906 glomérulos individuais. Os modelos de CNN YOLOv4 e VGG-16 foram empregados para localizar os glomérulos e classificar as lesões glomerulares em múltiplas classes relacionadas ao grau de severidade da lesão. De maneira geral os autores relataram uma precisão de 95,1% na classificação de leves/graves ou lesões escleróticas.

[Santos et al. \(2019\)](#) desenvolveram um método de classificação de imagens de glomérulo com doença de lesões mínimas (DLM) ou Glomeruloesclerose segmentar e focal (GESF). A base de imagens utilizada contém 87 imagens, sendo 33 de GESF e 54 de DLM. A proposta utiliza técnicas de transferência de aprendizado no formato de extração de características de 3 CNNs pré-treinadas. A saída é produzida por um classificador *Random Forest* (RF), com o qual alcançou acurácia de 93,3%.

Em um trabalho posterior, ([SANTOS et al., 2021](#)) expandiram a quantidade de imagens em 20%, tendo inserido nos experimentos bases de outras fontes ainda balanceou as classes e aplicou aumento de dados. Outro aspecto é que além de ter gerado mais *deep-features* foram também avaliadas características de textura de Haralick ([HARALICK; SHANMUGAM et al., 1973](#)) e geoestatísticas ([ISAACS; SRIVASTAVA; \(FIRM\), 1989](#)). A acurácia foi menor que no trabalho anterior (90,3%). Outro ponto relevante dos resultados é que os estudos apontaram que as características de textura são muito relevantes para a classificação: um filtro de atributos baseado na informação mútua selecionou muito mais características de textura que *deep-features* ao vetor de descrição das imagens.

Já a proposta de [Barros et al. \(2022\)](#) empregou um modelo de rede neural

profunda para identificar podocitopatia no glomérulo. A base de dados utilizada contém 835 imagens tratadas com PAS, H&E, Tricrômico e PAM. Em sua abordagem, usaram técnicas de transferência de aprendizado, usando a CNN VGG-19 como um extrator de características, seguindo para a etapa de classificação (realizada por meio de uma arquitetura de rede neural densa). O método proposto obteve resultados na classificação com 88,9% de precisão e 93,2% de recall.

Alguns trabalhos encontrados na literatura se aproximam ainda mais da presente proposta (ARAÚJO et al., 2017; MARSH et al., 2018; KANNAN et al., 2019; PESCE et al., 2022), pois além de lidar com imagens de citologia do glomérulo, também apresentaram propostas para detecção de glomeruloesclerose.

Araújo et al. (2017) trabalharam em um método para detectar glomeruloesclerose segmentar. O conjunto de dados utilizado no estudo era composto de 747 imagens de glomérulos individuais tratados com os corantes H&E e PAS. Neste trabalho os autores avaliaram diferentes formas de caracterizar a imagem (Gabor, Histograma e outros descritores de textura) em seguida forneceram esses vetores para quatro classificadores binários KNN, máquina de vetor de suporte (*support vector machine* - SVM), Naive Bayes, e Rede Neural Artificial (RNA). Os resultados apontaram que a classificação realizada com a rede neural foi mais eficiente alcançando 84,8% de acurácia nas imagens tratadas com H&E e 81,3% com o corante PAS.

No trabalho de Marsh et al. (2018) foi apresentado o desenvolvimento de um modelo de aprendizado profundo baseado na CNN VGG-16 que identifica e classifica os glomérulos com ou sem esclerose e regiões sem glomérulos. A base de imagens foi construída a partir de 48 imagens de lâminas inteiras de biópsia de rim tratadas com H&E, de onde foram separados 3.867 glomérulos individuais e mais 1.932 imagens de regiões sem glomérulos. Os autores relataram que o modelo obteve uma precisão de 89,3% e sensibilidade de 86,5% na categorização de glomérulos com esclerose.

Já no trabalho de Kannan et al. (2019) o conjunto de dados foi produzido a partir de 275 imagens de lâmina inteira coradas com tricrômico. Por meio de uma operação de janela deslizante, foram extraídas 745 imagens de glomérulos e 751 imagens que não continham glomérulos. A proposta dos autores é uma CNN capaz de classificar porções de lâminas em três classes: sem glomérulo, glomérulo normal ou parcialmente esclerosado (NPS) e glomérulo com esclerose (global). Segundo os autores, o modelo CNN pode discriminar com precisão imagens não glomerulares de imagens NPS e GS com uma acurácia de 92,67%.

Pesce et al. (2022) propuseram um sistema CAD para discriminar glomérulos com ou sem esclerose. No estudo eles compararam duas abordagens baseadas em modelos de Redes Neurais Artificiais. A primeira se baseou em um *ensemble* de RNA rasa que tinha como entrada características manualmente definidas. A outra empregou o *IBM*

Watson Visual Recognition System - um modelo baseado em rede neural profunda que recebe a imagem como entrada diretamente. O conjunto de dados utilizado consiste de 428 imagens de glomérulos com esclerose e 2.344 glomérulos sem esclerose; todas as amostras foram tratadas com o corante PAS. Nos resultados utilizando a primeira abordagem, a média de precisão e recall foi 98,4% e 93,1%, respectivamente; já com o IBM Watson foi de 96,5% e 94,3%.

A Tabela 6 apresenta um sumário dos trabalhos relacionados destacados.

Tabela 6 – Sumário dos trabalhos relacionados.

Trabalho	Dataset ¹	Corantes	Patologia	Resultados ²	OBS/ Técnicas
Barros et al. (2017)	811 63%	H&E, PAS	PGL	Acc: 85%	KNN
Ginley et al. (2018)	514 45%	PAS	Estágios de ND	Espec: 93%	Naive Bayes
Chagas et al. (2020)	811 63%	H&E, PAS	PGL e subcategorias	Prec: 82%	<i>Deep-features</i> , CNN, SVM
Zheng et al. (2021b)	3.906	PAS	Nefrite lúpica	Prec: 95,1%	CNNs YOLO e VGG
Santos et al. (2019)	87 38%	H&E, PAS, Pratas e Tricrômico	DLM, GESF	Acc: 93,3%	<i>Deep-features</i> , CNNs e RF
Santos et al. (2021)	104 50%	H&E, PAS, Pratas e Tricrômico	DLM, GESF	Acc: 90,3%	<i>Deep-features</i> , CNNs, Textura, Aumento de dados e RF
Barros et al. (2022)	835 41%	PAS, H&E, PAM, Tricrômico	Podocitopatia	Prec: 88,9%; Recall: 93,2%	<i>Transfer-learning</i> , CNN VGG-19, rede neural profunda
Araújo et al. (2017)	747 56%	PAS, H&E	GS	Acc: 84,8% (H&E); 81,3% (PAS)	Redes Neurais
Marsh et al. (2018)	3.867 22%	H&E	GS	Prec: 89,3%; sensib: 86,5%	<i>Deep learning</i> , CNN VGG-16
Kannan et al. (2019)	745 18%	Tricrômico	GS	Acc: 92,7%	CNN Inception, <i>transfer learning</i> , aumento de dados
Pesce et al. (2022)	2.772 15%	PAS	GS	Prec: 98,4%; Recall: 94,3%	IBM Watson, Redes Neurais, <i>Ensemble</i>

1 - Na coluna Dataset há o total de imagens e a porcentagem de imagens da classe positiva, respectivamente (quando se trata de classificação binária).

2 - Nos resultados destacamos o melhor atingido ou aquele da tarefa de classificação mais similar à esta proposta, para os trabalhos com multi classe.

3.1 Considerações Finais

Um ponto em comum nos trabalhos presentes na literatura é a utilização de técnicas baseadas em *deep learning*, mais especificamente a avaliação de arquiteturas de CNNs. Os pesquisadores têm obtido resultados relevantes utilizando modelos de CNNs já consolidados, como DenseNet, ResNet, VGG e Inception. Isto aponta que apesar das variadas possibilidades de construir arquiteturas de CNNs para implementação de um sistema de auxílio ao diagnóstico, as arquiteturas consolidadas são as que tendem a apresentar melhor efetividade. No entanto, o uso de *ensembles* ainda carece de investigação.

Neste capítulo, foi feita uma revisão das produções científicas relevantes relacionadas a sistemas CAD de doenças renais. Além disso, foram feitas considerações sobre as abordagens apresentadas e as técnicas utilizadas, a fim de obter um panorama do que vem sendo produzido pela comunidade científica em relação a esse tema.

4 Materiais e métodos

Este capítulo traz informações sobre a base de imagens utilizada e sobre o conjunto de experimentos realizados para se chegar ao método proposto para detectar glomeruloesclerose em imagens de biópsia de rim.

A partir da análise de trabalhos do estado da arte selecionamos as tecnologias que foram empregadas com êxito no campo da classificação de imagens médicas. Tendo essa base, elaboramos o fluxo de trabalho (Figura 15) que se divide em quatro fases: preparação do dataset, modelagem de arquiteturas, treinamentos de modelos e formação de comitês.

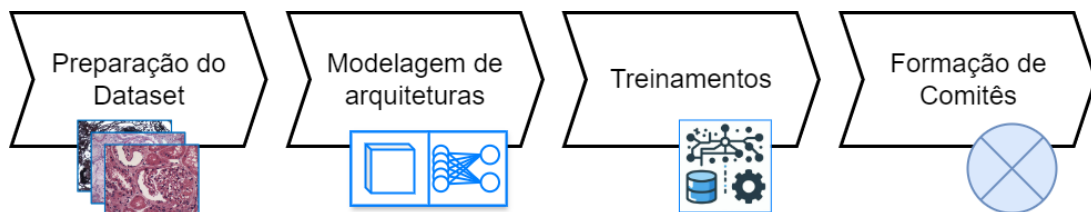


Figura 15 – Fluxograma dos experimentos.

Na preparação do dataset, procedemos com a revisão e organização de imagens de biópsia de rim de fontes diferentes de acordo a classe e corante aplicado na amostra. Aqui também definimos operações de pré-processamento a serem realizadas nas imagens antes de sua utilização em cenários específicos. A Seção 4.1 traz mais informações sobre essa fase.

Na fase de modelagem de arquiteturas, estão incluídas as atividades de análise e seleção de CNNs de referência em visão computacional. Também ocorre aqui a realização de modificações no design das arquiteturas para adaptá-las ao problema de classificação em questão no presente trabalho. Mais detalhes sobre essa fase estão presentes na Seção 4.2.

Na fase de treinamentos, as arquiteturas modeladas terão seus pesos ajustados para classificação binária da base de dados preparada na primeira fase. Duas estratégias de treinamento foram empregadas: na estratégia denominada *One-Step Tuning* (OST), a rede inteira é treinada em uma rodada. A outra estratégia divide o treinamento em duas etapas, *Shallow Fine-Tuning* (SFT) e *Deep Fine-Tuning* (DFT). Os detalhes da fase de treinamento encontram-se na Seção 4.3

Após todos os treinamentos das redes, seguimos para a formação de comitês. Neste ponto, os classificadores treinados passaram por análises de desempenho e apuração de métricas de classificação alcançadas em todas as etapas das estratégias

de treino. Com essa averiguação em mãos, definimos critérios de agrupamentos e selecionamos os melhores modelos para atuarem como membros votantes em comitês de classificação. Sendo o mecanismo de combinação parte indispensável de um comitê, aplicamos diferentes funções de apuração aos *ensembles* formados. Por fim, o desempenho obtido com os comitês é avaliado por meio de métricas de classificação. A explanação completa acerca dos comitês elaborados encontra-se na Seção 4.4.

4.1 Conjuntos de dados (dataset)

A base de dados utilizada foi formada a partir de imagens de três fontes. A maioria delas é do conjunto de dados (aqui denominado PSK2) construído por meio de colaboração de patologistas do Instituto Gonçalo Moniz da Fundação Oswaldo Cruz, e disponibilizado no âmbito do projeto PathoSpotter¹. As demais foram extraídas das bases de dados DME e INetDB encontradas no trabalho de Santos et al. (2021).

O conjunto de dados formado contém ao todo 2.158 imagens de microscopia óptica de biópsias renais, e evidenciam um glomérulo por imagem. Dessas imagens, 58,3% são diagnosticadas com glomeruloesclerose e os 41,7% restantes são saudáveis. As amostras de material biológico foram preparadas com os seguintes corantes químicos (*stains*): hematoxilina-eosina (H&E), ácido periódico de Schiff (PAS) e prata metenamina e de Jones (PMJ). A Tabela 7 mostra a distribuição da quantidade de imagens por corante e classe e base de origem.

Tabela 7 – Quantidade de imagens na base de dados.

Base	H&E	PAS	PMJ	Total
PSK2	526 642	232 397	141 203	899 1.242
DME	0 3	0 6	0 0	0 9
INetDB	0 1	0 7	0 0	0 8
Total	526 646	232 410	141 203	899 1.259

Cada par de valores indica a quantidade de imagens saudáveis e com glomeruloesclerose, respectivamente.

Na Figura 16 temos uma amostra de imagens das bases utilizadas. Verificamos que o tecido é complexo, apresenta coloração variada (principalmente devido aos corantes químicos utilizados no tratamento do material biológico), também há variedade no que se refere à texturas e formas dos glomérulos. Além disso, existem imagens visualmente semelhantes pertencentes a diferentes classes e vice-versa. Tais características tornam a tarefa de classificação mais desafiadora.

¹ <https://pathospotter.bahia.fiocruz.br>

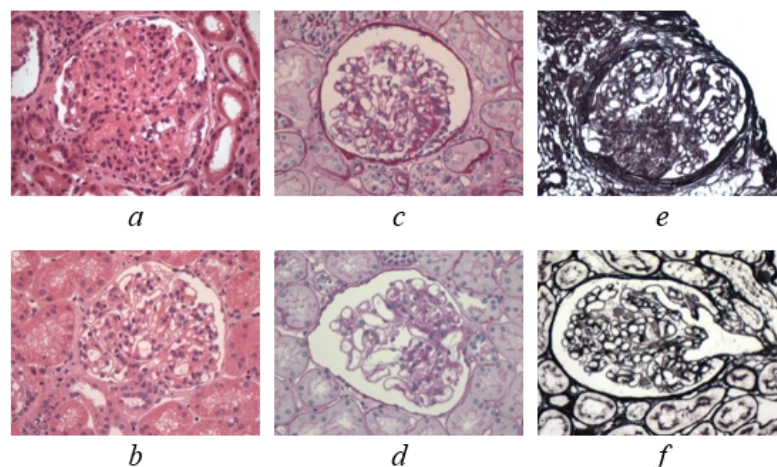


Figura 16 – Amostra de imagens das bases utilizadas. Corantes: *a* e *b*: H&E; *c* e *d*: PAS ; *e* e *f*: PMJ. A primeira linha contém glomérulos com esclerose e a segunda contém glomérulos saudáveis.

4.1.1 Validação Cruzada

Empregamos validação cruzada *k-fold* para avaliar o desempenho individual dos classificadores e dos *ensembles*. Essa técnica distribui aleatoriamente as instâncias do conjunto de dados em k subconjuntos mutuamente exclusivos (*folds*) de tamanhos aproximadamente iguais. A validação cruzada prevê que o algoritmo seja executado k vezes, assim a avaliação dos modelos é obtida pelo valor médio das métricas obtidas em cada execução.

Cada rodada usa um subconjunto diferente para teste e os $k - 1$ subconjuntos restantes são usados para desenvolvimento (treino e validação) dos modelos. Dessa forma, cada CNN é testada k vezes. Esse mecanismo garante que a performance do classificador seja avaliada em 100% dos elementos do conjunto de dados, ao fim de todas as rodadas do *k-fold*. A Figura 17 ilustra o mecanismo de funcionamento do *k-fold*.

Dividimos o conjunto de dados da imagem em cinco *folds* ($k = 5$). Portanto, em cada execução, 20% das instâncias compõem o subconjunto de teste – este não é utilizado durante o treinamento da rede, mas sim na avaliação do desempenho final dos classificadores. As imagens pertencentes aos *folds* restantes foram divididas em dois subconjuntos: o subconjunto de treinamento (70% do total) - os pesos da rede são ajustados com base na *loss* desse subconjunto. E o subconjunto de validação (10% do total) foi usado para monitorar a aprendizagem da rede durante seu treinamento.

Ainda tomamos o cuidado de fazer essa divisão de forma estratificada, desse modo preservamos em cada *fold* mesma proporção observada no conjunto de dados original, tanto no que se refere à classe quanto ao corante (Tabela 8). Isso impede que em uma determinada execução o modelo seja testado em um conjunto que eventualmente contenha imagens de apenas um corante ou de uma única classe - ou apresente um

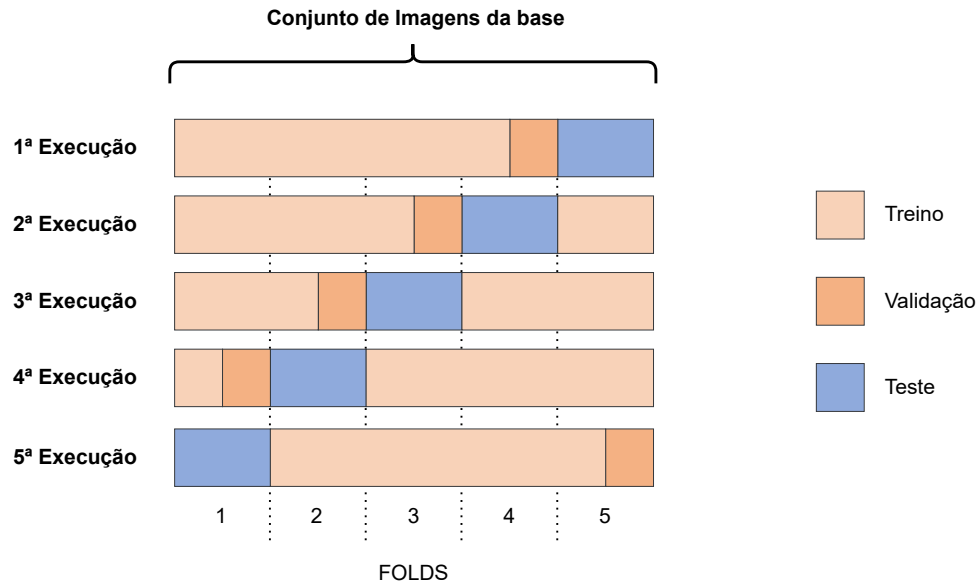


Figura 17 – Estratégia k -fold de validação cruzada para divisão do conjunto de dados. Com $k = 5$ temos *folds* de tamanho 20%. A distribuição das classes e corantes em cada *fold* é semelhante à observada na base inteira.

desbalanceamento muito acentuado nesse sentido. Tal cuidado contribui para que as métricas aferidas nos conjuntos de testes sejam mais confiáveis e reflitam o desempenho medido em toda a variedade de classes e corantes.

Tabela 8 – Quantidade de imagens nos subconjuntos de treino, validação e teste em uma execução do k -fold.

Subconjunto	H&E		PAS		PMJ		Total
Treino	368	452	163	287	99	142	630 881
Validação	53	65	23	41	14	20	90 126
Teste	105	129	46	82	28	41	179 252
Total	526	646	232	410	141	203	899 1.259

Cada par de valores indica a quantidade de imagens saudáveis e com glomeruloesclerose, respectivamente.

A proporção de imagens da classe positiva e negativa em cada conjunto e corante é a mesma observada na linha de totais (estratificação dos folds).

4.1.2 Aumento de dados e pré-processamento

As técnicas de aumento de dados consistem em gerar dados de forma artificial com a finalidade de ampliar a variedade de amostras disponíveis para treinamento e melhorar a capacidade de generalização dos modelos (SHORTEN; KHOSHGOFTAAAR, 2019). De acordo com Li et al. (2021), o aumento de dados é essencial em métodos baseados em aprendizado profundo, independentemente do problema.

O melhor desempenho na capacidade de generalização se deve também pela

redução do *overfitting*. Existem outras técnicas utilizadas para a mesma finalidade, a maioria dessas estratégias se concentram na própria arquitetura do modelo, como *batch normalization* ou *dropout*. Em contraste, o aumento de dados tenta solucionar o *overfitting* na origem do problema – o conjunto de dados de treinamento (CLARO et al., 2022). Gerar mais imagens para treino evita que o modelo dê atenção aos ruídos eventualmente presentes na imagem e os utilize para decidir a classe.

Considerando os benefícios do aumento de dados, aplicamos transformações às imagens de entrada no conjunto de treino gerando novas imagens artificialmente. É importante destacar que é necessário cautela quanto às transformações a serem realizadas. Operações como cisalhamento, oclusão, recortes ou zoom, podem causar efeitos indesejados para o tipo de imagem em questão, tais como, deformações da imagem, omissão de regiões com lesão, alteração de textura, formas e proporções de achados histológicos. Podendo, portanto, ocasionar a perda de características essenciais para a classificação.

Tendo isso em vista, definimos duas transformações espaciais simples: inversão vertical e inversão horizontal, e transformações sobre na intensidade das cores: adição ou redução de até 20% no brilho e contraste. Para o presente conjunto de dados, essas são transformações que não devem afetar a classificação da imagem (*label-preserving transformations*) (SHORTEN; KHOSHGOFTAR, 2019), além de refletirem situações que podem acontecer naturalmente no ambiente real quando o especialista analisa a amostra (ex. posicionamento da lâmina no microscópio, condições de iluminação e variações de qualidade dos corantes).

A aplicação dessas transformações acontece aleatoriamente nas imagens em tempo de execução do treinamento, antes de cada imagem ser fornecida para a rede em cada época de treinamento. Não há aplicação de aumento de dados nos conjuntos de testes para não prejudicar a confiabilidade das métricas aferidas. A Figura 18 exemplifica as operações de aumento de dados definidas.

As CNNs avaliadas neste trabalho possuem diferentes dimensões de entrada, quando comparadas às dimensões das imagens presentes na base de dados. Tal diferença é observada tanto em valores absolutos (dimensão) quanto em termos relativos (proporção). Sendo assim, mostra-se necessário realizar operações para adequação da imagem ao padrão de entrada das CNNs.

Para realizar essa adequação, adicionamos bordas pretas (preenchimento) às imagens para torná-las com o aspecto quadrado e depois a redimensionamos para o tamanho desejado (Figura 18a), por meio de interpolação bilinear. Esse procedimento evita deformações no formato dos objetos presentes na imagem. Uma simples alteração de proporção poderia esticar ou comprimir a textura e objetos de forma não uniforme e uma eventual perda de dados caso se realizasse um recorte quadrado de regiões

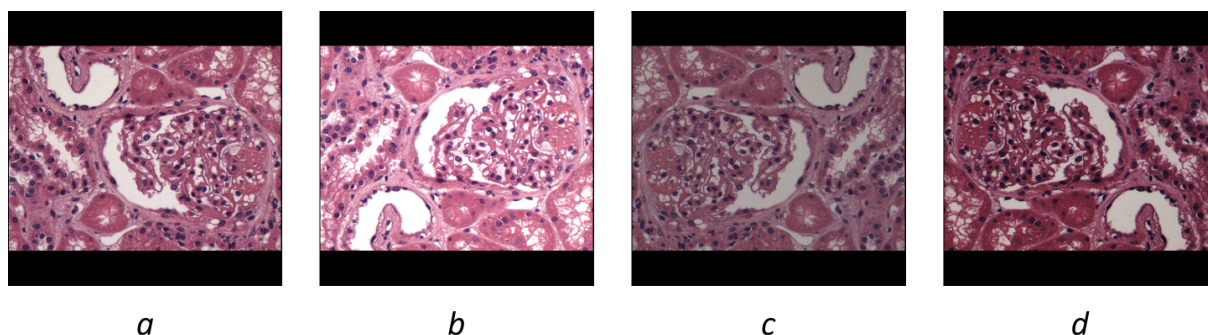


Figura 18 – Exemplos das operações de aumento de dados. Em *a* temos a imagem de entrada sem transformações, a partir dela são geradas as demais artificialmente. Em *b* foi aplicado inversão vertical e aumento de brilho; em *c*, inversão horizontal, redução de brilho e contraste; e em *d* inversão vertical e horizontal, aumento de contraste e redução de brilho.

específicas da imagem. Vale ressaltar que foram tomados cuidados para evitar a ocorrência de *aliasing* (GONZALEZ; WOODS, 2008).

Antes de serem levadas para as CNNs, as imagens ainda passam por operações de processamento específicas de cada arquitetura. Essas operações podem incluir o ajuste na escala de valores dos pixel, dimensionamento, reordenação de canais de cor e normalização. Dessa forma, as imagens adquirem a mesma representação para a qual a rede foi pré-treinada.

4.2 Modelagem de CNNs

Considerando os desafios de criar e treinar uma CNN do zero, como o alto custo computacional e a necessidade de grandes quantidades de dados, adotamos a estratégia de transferência de aprendizado (*Transfer Learning* - TL). Ao importar a arquitetura e os pesos de uma CNN pré-treinada na ImageNet, permitimos que nossa rede inicie o treinamento com um conhecimento prévio robusto sobre características visuais. Essa abordagem reduz o tempo de treinamento, e também melhora a capacidade do modelo de extrair características relevantes, mesmo com conjuntos de dados menores.

A ideia por trás do TL é aproveitar o conhecimento aprendido pela rede neural em um campo e aplicá-lo em outro campo relacionado (TAJBAKHSH et al., 2016). Vale ressaltar que reutilizamos, apenas a parte convolucional das CNNs, sendo a parte densa substituída por novas camadas totalmente conectadas, adaptadas à tarefa de classificação do presente trabalho. A Figura 19 representa esse procedimento de TL e os ajustes aplicados na modelagem das arquiteturas.

Criamos cinco alternativas para a parte densa, a mais simples contém apenas a camada de classificação com dois neurônios; as demais incluem ainda uma camada oculta com 64, 256, 512 ou 1.024 neurônios. A função de ativação na camada oculta é a

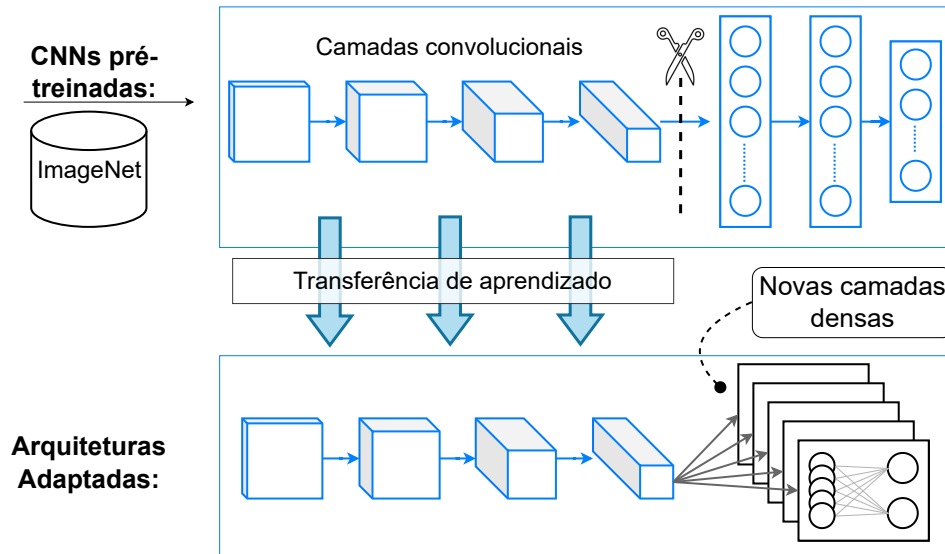


Figura 19 – Modelagem de arquiteturas. A parte convolucional de CNNs pré-treinadas na ImageNet é utilizada juntamente com seus pesos. A parte densa é substituída por novas camadas.

sigmoid e na camada de classificação a *softmax*, que produz como saída uma distribuição de probabilidade sobre ambas as classes. Inserimos uma camada de *Global Average Pooling* para fazer a ligação entre a parte convolucional e a nova parte densa. A Figura 20 evidencia as partes densas elaboradas.

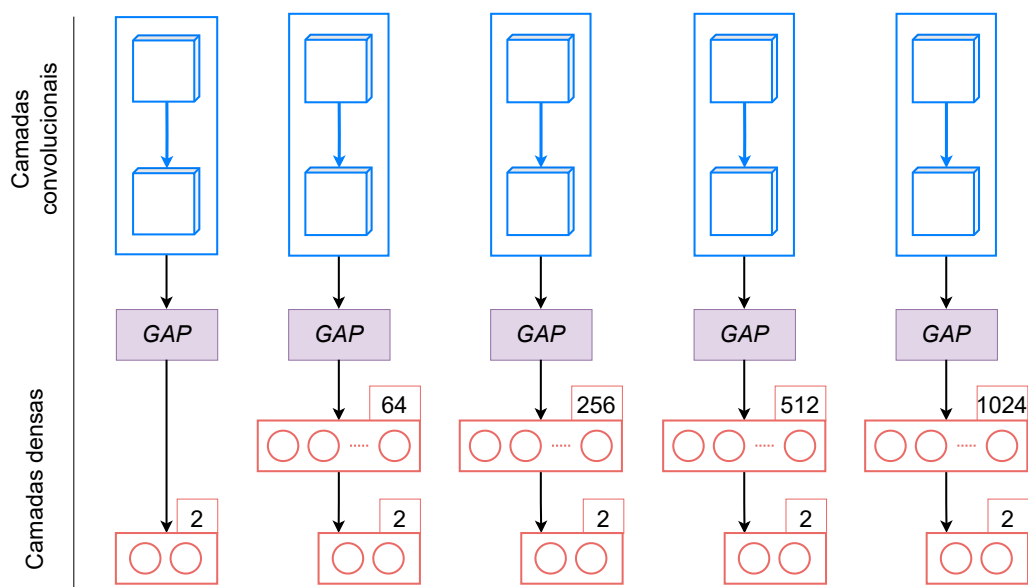


Figura 20 – Cinco variações de partes densas elaboradas. As camadas convolucionais das CNNs são conectadas às camadas densas após uma operação de *Global Average Pooling* (GAP). O rótulo numérico sobre as camadas densas indica a quantidade de neurônios na camada.

Ao fazer análise do estado da arte, identificamos arquiteturas disponíveis publicamente, com pesos já treinados na ImageNet e que foram empregadas com sucesso em tarefas de classificação semelhantes às do nosso trabalho. Com esse foco,

selecionamos CNNs bem estabelecidas na comunidade científica e largamente utilizadas em trabalhos relacionados.

Tendo em vista que a variedade é um critério importante para a escolha dos membros de um comitê, também buscamos promover diversidade nas topologias na seleção de CNNs, incluímos topologias mais tradicionais (do tipo sequencial), também com fluxos residuais, e outras que empregam abordagens como técnicas de regularização ou convolução dilatada.

Levando em consideração esses princípios selecionamos cinco arquiteturas de CNNs: DenseNet-201 (HUANG; LIU; WEINBERGER, 2016), EfficientNet-B2 (TAN; LE, 2019), Inception-V3 (SZEGEDY et al., 2016), ResNet-50 (HE et al., 2015) e VGG-19 (SIMONYAN; ZISSERMAN, 2015). A Tabela 9 apresenta algumas características dessas redes.

Tabela 9 – Resumo das arquiteturas de CNNs avaliadas.

Nome	Abbr.	Parâmetros (original)	Parâmetros (sem parte densa)	Top-5 acc (ImageNet)
DenseNet-201	dsn	20.242.984	18.321.984	93,6%
EfficientNet-B2	efn	9.177.569	7.768.569	94,9%
Inception-V3	icp	23.851.784	21.802.784	93,7%
ResNet-50	rsn	25.636.712	23.587.712	92,1%
VGG-19	vgg	143.667.240	20.024.384	90,0%

- Top-5 acc: desempenho do modelo no conjunto de validação do dataset ImageNet.

Outro ponto de destaque nessa seleção, é o desempenho obtido com as arquiteturas na base de dados ImageNet, onde se observa acurácia da ordem de 90% ou mais. Isso aponta para a grande capacidade de generalização dessas CNNs, e de acordo com Kornblith, Shlens e Le (2019), quanto melhor for o desempenho da arquitetura no conjunto de dados ImageNet, melhor será a transferência para outros conjuntos de dados de imagens naturais.

Para nos referirmos a uma arquitetura específica de forma mais prática, utilizaremos as abreviações da Tabela 9, seguidas da indicação de sua parte densa, separadas por um caractere underscore “_”. Ex. *rsn_64* representa a ResNet-50 com uma camada oculta de 64 neurônios seguida da camada de saída com dois neurônios; *dsn_2*: DenseNet-201 sem camada oculta, apenas a camada de saída.

Ao todo, foram produzidas 25 arquiteturas (considerando o produto de 5 CNNs-base e 5 partes densas). Ainda fizemos alguns testes com outras variações densas, mudando a quantidade de neurônios e de camadas ocultas, mas os resultados preliminares não apontaram ganhos de performance de classificação com as arquiteturas mais profundas.

4.3 Estratégias de treinamento

Durante o treino, um modelo CNN é alimentado com os dados de entrada, e seus parâmetros internos (pesos) são ajustados para minimizar o erro (calculado pela função de *loss*) entre suas predições e as classes corretas (aprendizagem supervisionada). É importante destacar que as redes foram pré-treinadas sobre uma base de dados distinta e para uma tarefa de classificação diferente da que precisamos. Além disso a parte densa foi substituída e portanto não está pré-treinada.

Diante disso, seguimos duas estratégias de treinamento. A *One-Step Tuning* (OST) segue o caminho tradicional, onde a rede inteira é treinada sobre a base de dados. A outra estratégia divide o treinamento em duas etapas, a *Shallow Fine-Tuning* (SFT) e a *Deep Fine-Tuning* (DFT).

Na etapa SFT, os pesos das camadas convolucionais são congelados (fixados) e apenas os pesos das camadas densas, que foram adicionadas depois, são ajustados durante o treinamento. Depois, na etapa DFT, toda a rede é treinada com uma taxa de aprendizagem bem baixa. No entanto, na DFT o treinamento não parte do estado inicial da rede; aqui o ponto de partida é o modelo já produzido na SFT. A DFT pode ser vista como uma continuação do treinamento da SFT. A Figura 21 apresenta as duas estratégias mencionadas.

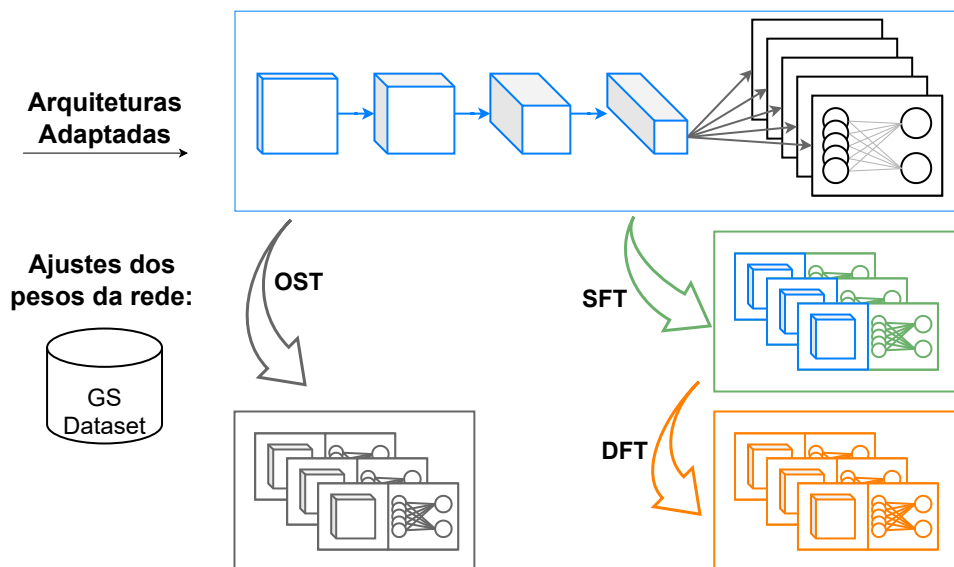


Figura 21 – Estratégias de treinamentos aplicadas. Em um estágio: Todas as camadas da rede são treinadas juntas desde o início (OST). Em dois estágios: Primeiro são treinadas as camadas densas (SFT), depois toda a rede é refinada (DFT).

A Tabela 10 apresenta detalhes dos parâmetros de treinamento utilizados. O processo de treinamento da rede foi acompanhado através do valor da função de perda (*loss*) durante 80 épocas no estágio SFT, conforme ilustrado na Figura 22. O estado da

rede foi salvo na época em que ocorreu o menor valor de perda no conjunto de validação. Em seguida, a etapa DFT recupera esse modelo, salvo na SFT, e continua o treinamento a partir desse ponto. A DFT é executada por mais 80 épocas ou é interrompida se não houver melhorias na *loss* por 20 épocas consecutivas, implementando assim uma estratégia de parada antecipada.

Tabela 10 – Parâmetros utilizados durante o treinamento das CNNs.

Parâmetro	Valor
Otimizador	RMSprop (TIELEMAN; HINTON et al., 2012)
Taxa de aprendizagem	OST e SFT: 10^{-3} ; DFT: 10^{-6}
Função de <i>loss</i>	Categorical crossentropy
Número de épocas	OST e SFT: 80; DFT: de 21 até 80 (parada antecipada)

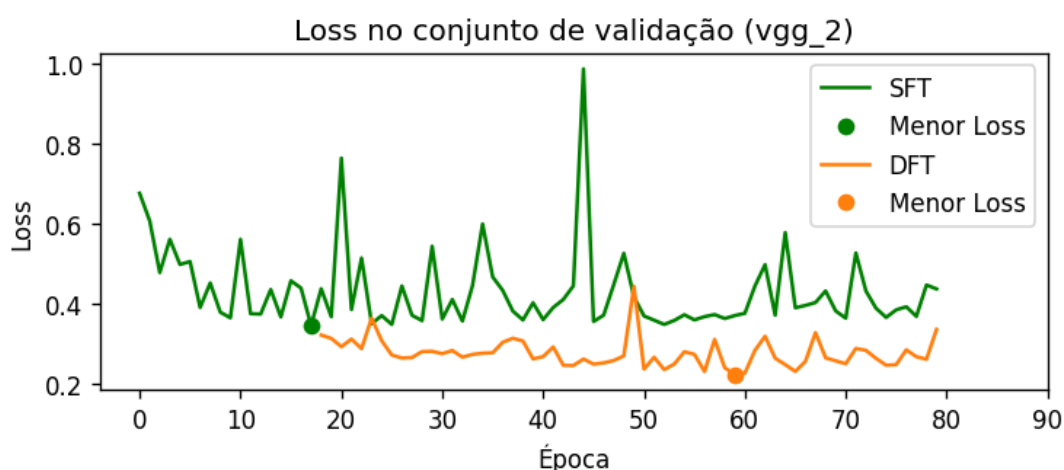


Figura 22 – Monitoramento da *loss* no conjunto de validação durante o processo de treinamento. Os pontos destacados mostram a época de menor *loss*. A etapa DFT continua o treinamento a partir do modelo salvo na época de menor *loss* da SFT.

4.4 Formação de Comitês

Após todos os treinamentos das redes, dispomos dos classificadores base. No entanto, para formar os comitês de classificação é preciso ainda selecionar aqueles que serão membros nos *ensembles*. É fundamental que cada membro de um comitê possua um bom desempenho individualmente. Com este foco, para selecionar os melhores membros, comparamos a performance dos classificadores através da métrica AUC, uma vez que essa métrica avalia a capacidade do modelo em separar as classes, considerando vários limiares (v. Seção 2.6). Vale ressaltar que para essa finalidade consideramos a AUC aferida no conjunto de validação.

Outro aspecto importante que buscamos na formação dos comitês foi a promoção da variedade dos membros votantes. Para tanto, não consideramos apenas o desempenho

dos classificadores, mas definimos critérios específicos de agrupamento dos membros dos comitês. Tais critérios implicam na formação de comitês que garantem a presença de arquiteturas distintas e também de CNNs em diferentes etapas de treinamento. O diagrama da Figura 23 representa o processo de formação dos comitês.

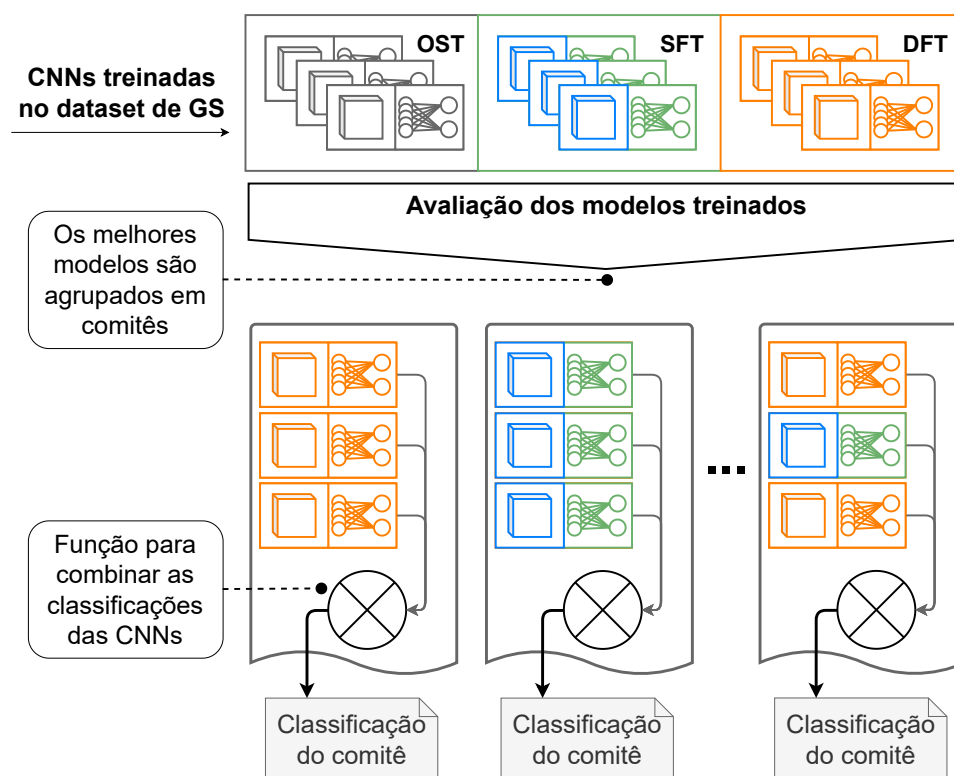


Figura 23 – Esquema de formação de comitês. Os modelos de melhor AUC no conjunto de validação são selecionados e formam grupos de acordo com critérios pré definidos. As saídas produzidas pelos membros são combinadas por funções de apuração, que produzem a classificação do comitê.

Os critérios de seleção dos classificadores votantes nos comitês levam em consideração aspectos referentes à quantidade de membros, estágios de treinamento e obrigatoriedade de CNNs de base distintas (diversidade). Combinando as possibilidades, há potencialmente 16 grupos. Conforme detalhado nos critérios a seguir:

- Quanto à quantidade de membros: **3** ou **5**.
- Quanto ao estágio de treinamento:
 - **O**: Apenas modelos no estágio OST;
 - **S**: Apenas modelos no estágio SFT;
 - **D**: Apenas modelos no estágio DFT;
 - **A**: Aberto a qualquer etapa/estratégia de treinamento.
- Quanto à diversidade de CNNs:

- **d**: Obriga que todas as redes presentes no comitê sejam de bases distintas;
- **r**: Permite a seleção de CNNs de mesma base.

Cada grupo produzido é nomeado de acordo com os critérios que o formaram (v. Figura 24). Por exemplo, **3Sr** é o grupo formado pelos 3 melhores classificadores no estágio SFT, sendo permitida repetições, i.e. é possível selecionar arquiteturas oriundas da mesma CNN-base; **5Dd** contém o melhor classificador de cada CNN-base (5 membros de bases distintas) no estágio DFT; **5Ar** é formado pelos cinco melhores classificadores, independente de qual estágio de treinamento e de que CNN-base sejam.

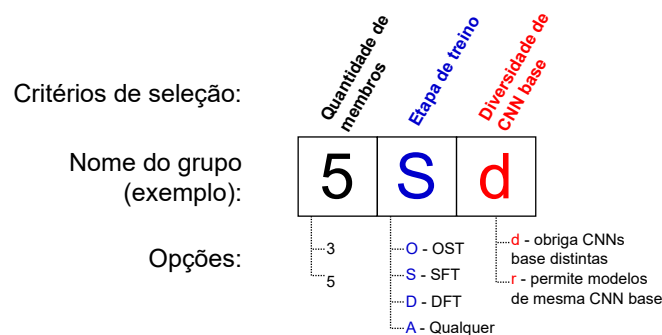


Figura 24 – Nomenclatura dos grupos formados nos comitês. No exemplo, o grupo 5Sd é formado pelo melhor classificador de cada CNN-base no estágio de treinamento SFT.

Tendo resolvido a seleção dos membros dos comitês, resta definir o mecanismo de combinação das predições (funções de apuração) dos classificadores para que o comitê seja capaz de produzir uma saída consolidada. Um mecanismo bem intuitivo e direto, conhecido como votação simples ou majoritária, consiste em checar a classe de saída de cada membro e identificar qual classe foi a mais votada, ela será o resultado do comitê.

O mecanismo de votação simples dá a mesma importância para todos os classificadores, independentemente do quão bom cada um seja. Outro dado que é subutilizado é a distribuição de probabilidade produzida na saída dos modelos (ativação *softmax*). Não faz diferença, no mecanismo de votação simples, se a classificação de saída teve 51% de probabilidade ou 99%, apenas é levado em consideração qual foi a classe mais provável.

Observando tais aspectos, elaboramos também outras funções de apuração, que levam em conta a distribuição de probabilidade (tipos P_2 , P_1 e P_0) e/ou ponderam a votação de um classificador baseado em medidas que refletem o quão boas ou confiáveis são suas classificações (tipos W_0 , W_1 e W_2). Em todas essas funções, a saída do comitê (CMT_{out}) é um par ordenado de valores numéricos normalizados, cuja soma é 100%, gerado conforme a Equação 4.1. O primeiro valor do par ordenado está associado à classe saudável e o segundo à classe GS. A classificação final é positiva caso o valor da

classe saudável seja menor que o valor da classe GS, caso contrário a classe de saída é negativa:

$$CMT_{out} = \left(\frac{H}{H+G} \quad , \quad \frac{G}{H+G} \right), \quad (4.1)$$

sendo:

H : Somatório vinculado às classificações negativas;

G : Somatório vinculado às classificações positivas.

A Equação 4.1 é baseada nos parâmetros H e G , seus valores são computados de formas diferentes para cada função de apuração elaborada. Vamos iniciar descrevendo as funções que não aplicam pesos às saídas produzidas pelos membros votantes (indicadas por W_0).

Sejam S_i^- e S_i^+ as probabilidades de saída (*softmax*) para as classes negativa e positiva, respectivamente, do i -ésimo classificador votante de um comitê formado por c classificadores. Na função de apuração denominada P_2W_0 , os valores de H e G são calculados pelas equações 4.2 e 4.3. Esse mecanismo considera as probabilidades de ambas as classes e faz o somatório das probabilidades geradas pelos classificadores para cada classe separadamente:

$$H = \sum_{i=0}^c S_i^- \quad (4.2)$$

e

$$G = \sum_{i=0}^c S_i^+ \quad . \quad (4.3)$$

As funções de apuração do tipo P_1 , consideram no cálculo do somatório apenas a probabilidade da classe de saída, ou seja, a de maior probabilidade. Para isso usamos as equações 4.4 e 4.5 que promovem essa operação anulando o valor da classe minoritária:

$$\alpha_i^- = \begin{cases} S_i^- & \text{se } S_i^- \geq S_i^+ \\ 0 & \text{se } S_i^- < S_i^+ \end{cases} \quad (4.4)$$

e

$$\alpha_i^+ = \begin{cases} S_i^+ & \text{se } S_i^+ > S_i^- \\ 0 & \text{se } S_i^+ \leq S_i^- \end{cases} \quad . \quad (4.5)$$

Logo, no mecanismo P_1W_0 , os parâmetros H e G são obtidos de acordo com as equações 4.6 e 4.7:

$$H = \sum_{i=0}^c \alpha_i^- \quad (4.6)$$

e

$$G = \sum_{i=0}^c \alpha_i^+ \quad . \quad (4.7)$$

Já os mecanismos que desconsideram o valor específico de probabilidade de saída do *softmax* (tipo P_0), categorizam a saída considerando o valor 1 para a classe de maior probabilidade e 0 para a outra. As equações 4.8 e 4.9 representam essa operação:

$$\beta_i^- = \begin{cases} 1 & \text{se } S_i^- \geq S_i^+ \\ 0 & \text{se } S_i^- < S_i^+ \end{cases} \quad (4.8)$$

e

$$\beta_i^+ = \begin{cases} 1 & \text{se } S_i^+ > S_i^- \\ 0 & \text{se } S_i^+ \leq S_i^- \end{cases} \quad . \quad (4.9)$$

A função de apuração P_0W_0 produz a mesma classificação final que o mecanismo de votação simples já explanado. Os parâmetros H e G dessa função são obtidos pelas equações 4.10 e 4.11:

$$H = \sum_{i=0}^c \beta_i^- \quad (4.10)$$

e

$$G = \sum_{i=0}^c \beta_i^+ \quad . \quad (4.11)$$

Na Figura 25 exemplificamos a aplicação das três funções de apuração apresentadas até o momento, para o caso de um comitê com cinco classificadores membros votantes. Nota-se que no mecanismo de votação simples (P_0W_0) a classificação de saída é negativa, já com as outras duas funções, o resultado é positivo, e apresentam valores probabilísticos diferentes.

Esses três mecanismos de combinação não levam em consideração nenhuma ponderação para dar mais ou menos importância à predição de um determinado classificador. Portanto, todos os membros votantes são tratados de maneira igualitária,

	P₂W₀	P₁W₀	P₀W₀
	(Saída softmax) Negativo Positivo	Negativo Positivo	Negativo Positivo
Classificador A	0,7 0,3	0,7 0	1 0
Classificador B	0,55 0,45	0,55 0	1 0
Classificador C	0,1 0,9	0 0,9	0 1
Classificador D	0,55 0,45	0,55 0	1 0
Classificador E	0,05 0,95	0 0,95	0 1
Somatório:	1,95 3,05	1,8 1,85	3 2
Saída Normalizada:	0,39 0,61	0,49 0,51	0,6 0,4

Figura 25 – Exemplo de aplicação das funções de apuração sem pesos (W_0). A borda mais espessa indica a classe de maior probabilidade (classe de saída). A distribuição de probabilidade é somada separadamente por classe e ao final há uma normalização. P_2W_0 calcula com ambas as probabilidades de cada saída. P_1W_0 só utiliza a o valor da classe de maior probabilidade de cada resultado. Já P_0W_0 não calcula com o valor da distribuição de probabilidade, apenas conta a classe de saída, equivalendo à votação simples.

ainda que alguns possam ter apresentado taxas de acerto maiores. Diante disso, implementamos os mecanismos que aplicam um peso para cada classificador (indicados por W_1).

Quanto ao uso da distribuição de probabilidades, de maneira semelhante, aqui também temos 3 métodos: P_2W_1 , P_1W_1 , P_0W_1 . A diferença desses mecanismos em relação aos anteriores (do tipo W_0) é que, antes de efetuar o somatório, as probabilidades consideradas em cada caso são multiplicadas pela acurácia de validação do respectivo classificador. Ao multiplicar a distribuição de probabilidade pela acurácia, inserimos um fator que indica o quanto podemos confiar na classificação daquela CNN, e assim ajustar proporcionalmente o seu valor de saída do *softmax*.

Seja acc_i a acurácia do i -ésimo classificador do comitê, medida no conjunto de validação. As equações 4.12 e 4.13 apresentam a forma de obtenção dos parâmetros H e G para a função de apuração P_2W_1 :

$$H = \sum_{i=0}^c (S_i^- \times acc_i) \quad (4.12)$$

e

$$G = \sum_{i=0}^c (S_i^+ \times acc_i) \quad . \quad (4.13)$$

Já no caso do mecanismo P_1W_1 , leva-se em conta a probabilidade de saída da classe majoritária em cada classificador e desconsidera-se a da classe minoritária. Nesta função de apuração os parâmetros H e G são obtidos conforme as equações 4.14 e 4.15 a seguir:

$$H = \sum_{i=0}^c (\alpha_i^- \times acc_i) \quad (4.14)$$

e

$$G = \sum_{i=0}^c (\alpha_i^+ \times acc_i) \quad . \quad (4.15)$$

No mecanismo P_0W_1 os valores específicos de probabilidade de saída não são considerados e os parâmetros H e G são aferidos conforme mostrado nas equações 4.16 e 4.17:

$$H = \sum_{i=0}^c (\beta_i^- \times acc_i) \quad (4.16)$$

e

$$G = \sum_{i=0}^c (\beta_i^+ \times acc_i) \quad . \quad (4.17)$$

A Figura 26 mostra o funcionamento desses métodos ponderados com um peso por classificador (W_1). Observa-se que neste exemplo, apesar das saídas originais dos classificadores (*softmax*) serem as mesmas dos exemplos W_0 (Figura 25), devido a aplicação de pesos os resultados probabilísticos se diferenciaram, com destaque ao P_1W_1 cuja classificação de saída mudou em relação ao exemplo anterior.

Os mecanismos tipo W_1 aplicam um único peso para cada CNN: a acurácia, que mede a taxa de acerto geral do classificador para ambas as classes indiscriminadamente. É possível ser mais preciso utilizando pesos de ponderação específicos a depender se a classe predita foi negativa ou positiva. As métricas npv e ppv (de validação) são as ideais para esse caso, pois representam respectivamente a taxa de acerto considerando cada uma dessas classificações. Nesse sentido elaboramos as funções que aplicam pesos diferentes a depender da classe (indicados por W_2).

Sejam npv_i e ppv_i respectivamente as métricas npv e ppv aferidas no conjunto de validação do i -ésimo classificador do comitê. Os parâmetros H e G para a função de apuração P_2W_2 são calculados conforme as equações 4.18 e 4.19 a seguir:

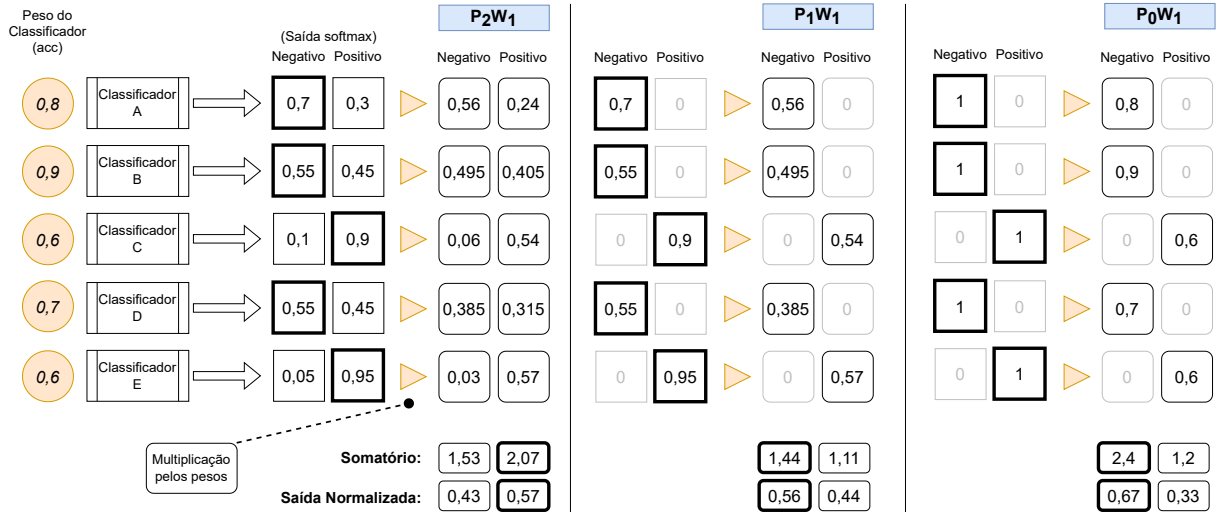


Figura 26 – Exemplo de aplicação das funções de apuração com um peso por classificador (W_1). Antes do somatório, faz-se a multiplicação das probabilidades de saída consideradas, pela acc do respectivo classificador.

$$H = \sum_{i=0}^c (S_i^- \times npv_i) \quad (4.18)$$

e

$$G = \sum_{i=0}^c (S_i^+ \times ppv_i) \quad (4.19)$$

As equações 4.20 e 4.21, demonstram a forma de cálculo dos parâmetros H e G para a função de apuração P_1W_2 :

$$H = \sum_{i=0}^c (\alpha_i^- \times npv_i) \quad (4.20)$$

e

$$G = \sum_{i=0}^c (\alpha_i^+ \times ppv_i) \quad (4.21)$$

Por fim, a função de apuração P_0W_2 obtém os parâmetros H e G de acordo com as equações 4.22 e 4.23 a seguir:

$$H = \sum_{i=0}^c (\beta_i^- \times npv_i) \quad (4.22)$$

e

$$G = \sum_{i=0}^c (\beta_i^+ \times ppv_i) \quad (4.23)$$

A Figura 27 mostra o funcionamento das funções de apuração implementadas que aplicam pesos diferentes para cada classe.

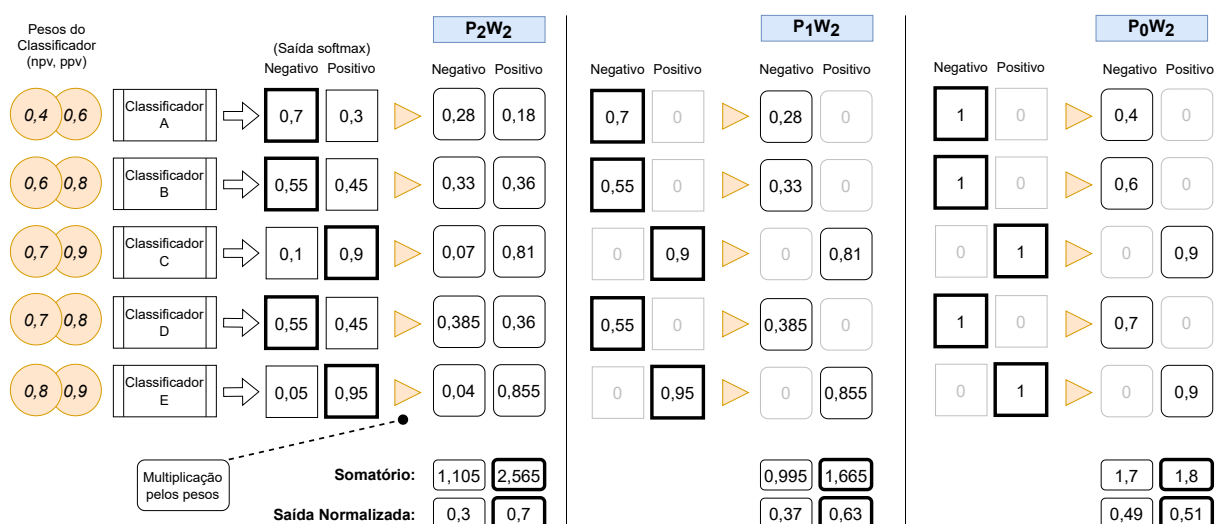


Figura 27 – Exemplo de aplicação das funções de apuração com um peso para cada classe (W_2). Difere do mecanismo W_1 , pois em vez de um peso único, usa o npv e ppv como pesos aplicados sobre a probabilidade para as classes negativa e positiva, respectivamente.

4.5 Considerações Finais

Esse capítulo detalhou as etapas seguidas nesta proposta para o desenvolvimento de método para classificar imagens de glomérulos com ou sem glomeruloesclerose. Entre os principais aspectos destacamos os seguintes pontos:

- Divisão da base via *k-fold* estratificado por corante e classe;
- Aumento de dados de treinamento com transformações *label-preserving* naturais;
- Transferência de aprendizado de 5 CNNs-base e 25 arquiteturas avaliadas ao todo;
- Estratégias de treinamento em estágio único (OST) e em dois estágios (SFT e DFT);
- Acompanhamento do treino com *loss* monitorada para prevenir *overfitting*;
- A criação de *ensembles* de classificadores agrupados por critérios bem definidos;
- Definição de 9 funções de apuração, incluindo opções probabilísticas e ponderadas.

Destacamos alguns recursos de software e hardware utilizados no desenvolvimento desse trabalho: linguagem de programação Python², bibliotecas relacionadas ao projeto SciPy (JONES et al., 2001) em especial scikit-learn (PEDREGOSA et al., 2011), e a

² <<https://www.python.org/>>

biblioteca Keras (CHOLLET et al., 2015) com *backend* TensorFlow (ABADI et al., 2015). A infraestrutura em nuvem Google Colaboratory³ foi utilizada como ambiente de execução. No próximo capítulo, serão apresentados os resultados obtidos.

³ <<https://colab.research.google.com/>>

5 Resultados e discussão

Neste capítulo detalhamos os resultados individuais obtidos com as 25 arquiteturas de CNNs avaliadas na Seção 5.1 e na seção seguinte os resultados alcançados utilizando os *ensembles*. Todas as métricas apresentadas foram obtidas a partir da média aritmética dos valores medidos nas 5 execuções da validação cruzada (*k-fold*).

5.1 Resultados com CNNs individuais

Dividimos esta seção em duas partes, de acordo com a estratégia de treinamento aplicada. Na Subseção 5.1.1, discorremos sobre os resultados obtidos com a estratégia OST e na Subseção 5.1.2, apresentamos os resultados com a estratégia de treinamento em dois estágios (SFT e DFT).

5.1.1 Resultados com treinamento em etapa única (OST)

Os resultados obtidos por meio da estratégia de treinamento em etapa única, com a maioria das arquiteturas apontaram um desempenho insatisfatório. A Tabela 11 apresenta as métricas obtidas por dez modelos em ordem decrescente de acurácia medida no conjunto de testes.

Tabela 11 – Resultados individuais com a estratégia OST.

Arquitetura	acc	kappa	ppv	npv	sensib	espec	AUC
dsn_2	87,6 ±0,8	74,5 ±1,8	88,8 ±2,0	86,1 ±1,6	90,2 ±1,6	84,0 ±3,6	94,4 ±0,6
rsn_2	84,2 ±1,6	67,2 ±3,5	84,8 ±3,1	83,9 ±3,6	89,2 ±3,3	77,3 ±5,9	91,9 ±1,5
efn_256	60,2 ±4,1	6,7 ±14,9	61,2 ±6,4	12,1 ±27,1	94,3 ±12,8	12,4 ±27,7	54,5 ±7,9
dsn_1024	59,5 ±2,6	3,3 ±7,5	59,1 ±1,7	18,7 ±41,7	99,8 ±0,4	3,1 ±7,0	52,0 ±3,5
efn_512	59,3 ±1,7	2,9 ±5,4	59,0 ±1,2	34,6 ±48,3	99,2 ±1,8	3,3 ±6,6	57,2 ±4,0
efn_2	59,2 ±2,0	2,5 ±5,5	58,9 ±1,3	19,0 ±42,6	99,9 ±0,2	2,2 ±5,0	56,9 ±9,6
vgg_256	59,2 ±1,9	2,5 ±5,5	58,9 ±1,2	18,3 ±40,8	99,8 ±0,4	2,3 ±5,2	50,1 ±3,6
dsn_512	59,1 ±1,7	3,1 ±7,0	59,2 ±2,0	11,9 ±26,5	97,2 ±6,2	5,7 ±12,7	51,5 ±2,5
efn_1024	58,5 ±0,5	0,5 ±1,3	58,4 ±0,4	16,7 ±37,3	99,8 ±0,2	0,6 ±1,2	55,6 ±3,3
rsn_256	58,3 ±0,1	0,0 ±0,0	58,3 ±0,1	0,0 ±0,0	100,0 ±0,0	0,0 ±0,0	49,4 ±0,8

– Valores em termos percentuais medidos nos conjuntos de testes. Os melhores resultados estão destacados em azul.

Apenas os dois primeiros modelos da tabela (dsn_2 e rsn_2) alcançaram desempenhos aceitáveis considerando os patamares observados em trabalhos semelhantes. O terceiro (efn_256) apresentou queda acentuada nas métricas (exceto na sensibilidade). Os demais modelos apresentados na Tabela 11 acompanharam gradativamente a mesma tendência. Os modelos não exibidos tiveram desempenho similar ao último (rsn_256).

À exceção de `dsn_2` e `rsn_2`, a acurácia dos modelos aproxima-se da própria proporção da classe majoritária (positiva) do dataset. O kappa, por já considerar essa distribuição em sua fórmula, se aproxima de zero indicando desacordo entre as predições dos modelos e as classificações do especialista. A especificidade encaminhou-se a zero mas a sensibilidade tomou caminho contrário, chegando a 100%. A análise dessas métricas indica que tais modelos estão tendenciosos a classificar todas (ou quase todas) as instâncias como positivas. Averiguações realizadas nas matrizes de confusão e confirmaram isso.

Estamos aplicando técnicas de transferência de aprendizado sobre as CNNs, de modo que os pesos das partes convolucionais são importados pré-treinados na ImageNet. Portanto, inicialmente eles estão ajustados para extrair características descritivas com eficiência. Porém, a parte densa é nova e inicializada com pesos aleatórios, então já nas primeiras épocas de um treinamento integral da rede, o erro (*loss*) gerado na saída é alto e impacta fortemente os pesos da parte convolucional devido a retro-propagação do erro (*backpropagation*). Isso compromete a capacidade de extração de características que fora aprendida com a ImageNet (TAJBAKHSH et al., 2016).

Devido a esses aspectos da estratégia de treinamento OST, a obtenção de resultados melhores com essas arquiteturas, demandaria por estudos, testes e avaliações mais aprofundadas, notadamente no que diz respeito a ajustes de hiperparâmetros de treinamento, outras variações de camadas densas e funções de ativação, e certamente, mais imagens e épocas para treinamento, haja visto que tais CNNs foram concebidas para o dataset ImageNet, que é cerca de 600 vezes maior do que o conjunto de dados utilizado aqui.

Por outro lado, vimos o treinamento em duas etapas como estratégia mais viável. Treinar apenas a parte densa primeiro corrige seus pesos (deixam de ser aleatórios) e adapta-a para tratar os mapas de características recebidos da parte convolucional. Após essa adaptação inicial, o erro (*loss*) gerado na saída já se torna menor, e então toda a arquitetura da rede pode ser treinada com o conjunto de dados específico, sem causar danos ao aprendizado, permitindo que a rede melhore seu desempenho no tipo de imagem presente na base utilizada (TAJBAKHSH et al., 2016).

5.1.2 Resultados com treinamento em duas etapas (SFT e DFT)

Apresentaremos nesta seção os métricas alcançadas por todas as arquiteturas nos estágios SFT e DFT medidos nos conjuntos de teste. Organizamos os resultados por CNN-base, dessa forma, inicialmente traremos os resultados dos modelos baseados na DenseNet-201, em seguida, os modelos EfficientNet-B2, Inception-V3, Resnet-50, e por último os da VGG-19.

Os resultados obtidos com as arquiteturas baseadas em DenseNet-201 estão dispostos na Tabela 12. Com a arquitetura adaptada com 256 neurônios na camada oculta (dsn_256), atingimos 90,6% de acurácia, kappa de 80,7%, sensibilidade e especificidade de 93,1% e 87,2%, ainda no estágio SFT. Ao comparar esses valores com os da Tabela 11, vemos que o melhor resultado com a mesma cnn-base (dsn_2) na estratégia OST teve desempenho inferior em todas as métricas.

Tabela 12 – Resultados das arquiteturas baseadas na CNN DenseNet-201.

Arquitetura		acc	kappa	ppv	npv	sensib	espec	AUC
SFT	dsn_2	89,9 ±1,4	79,1 ±2,9	90,6 ±1,4	89,0 ±3,1	92,3 ±2,6	86,5 ±2,4	96,2 ±0,3
	dsn_64	89,9 ±1,0	79,2 ±2,1	90,9 ±0,7	88,5 ±1,5	91,9 ±1,2	87,1 ±0,9	96,5 ±0,2
	dsn_256	90,6 ±0,4	80,7 ±1,0	91,1 ±1,1	90,0 ±1,1	93,1 ±0,9	87,2 ±1,9	96,7 ±0,3
	dsn_512	90,5 ±0,4	80,4 ±0,8	91,0 ±1,5	90,0 ±2,0	93,0 ±1,8	87,0 ±2,5	96,7 ±0,2
	dsn_1024	90,5 ±0,3	80,3 ±0,6	91,0 ±1,0	89,7 ±1,3	92,9 ±1,1	87,1 ±1,7	96,7 ±0,3
DFT	dsn_2	90,4 ±0,8	80,2 ±1,5	91,5 ±1,1	89,0 ±2,8	92,1 ±2,4	88,0 ±2,0	96,6 ±0,4
	dsn_64	90,6 ±0,8	80,6 ±1,6	91,7 ±1,4	89,0 ±1,2	92,2 ±1,0	88,3 ±2,2	96,8 ±0,4
	dsn_256	90,9 ±0,9	81,2 ±1,8	92,1 ±1,1	89,2 ±0,7	92,3 ±0,5	88,9 ±1,7	96,9 ±0,1
	dsn_512	90,7 ±0,7	80,9 ±1,5	91,7 ±1,1	89,4 ±1,1	92,5 ±0,9	88,2 ±1,6	97,0 ±0,1
	dsn_1024	90,6 ±0,7	80,7 ±1,5	91,5 ±1,2	89,5 ±0,4	92,6 ±0,4	87,9 ±1,8	97,0 ±0,1

– Valores em termos percentuais medidos nos conjuntos de testes.

– Os melhores resultados estão destacados em azul e os hachurados indicam que houve redução após DFT.

Comparando os estágios de treinamento (SFT e DFT) de cada arquitetura, verificamos que as métricas em geral apresentaram melhorias. O maior aumento foi na especificidade da dsn_256 que incrementou em 1,7 pp. Em contraponto, as métricas npv e sensibilidade reduziram na maioria das arquiteturas, a npv da dsn_256 foi a que mais reduziu – 0,9 pp. Essas duas métricas têm em comum a quantidade de FN no denominador, isso indica que um pequeno aumento na quantidade de falsos negativos pode ter ocasionado esse efeito.

Apresentamos na Tabela 13 os resultados alcançados utilizando as arquiteturas baseadas na EfficientNet-B2. Essa CNN obteve desempenho menor que as baseadas na dsn. Na etapa DFT, a arquitetura efn_64 alcançou 89,6% de acurácia, 78,6% de kappa, 91,3% de ppv, 87,9% de especificidade e 96,1% de AUC. O maior npv e sensibilidade foram alcançados com a efn_2, 88,0% e 91,6% respectivamente no estágio DFT.

Com exceção da sensibilidade da efn_256, que reduziu 0,1%, todas as demais métricas e arquiteturas apresentaram aumento com o treinamento no estágio DFT, destacando-se o kappa e a especificidade da efn_64 que incrementaram em 3,4 pp e 3,2 pp respectivamente.

A Tabela 14 apresenta os resultados observados nas arquiteturas baseadas na Inception-V3. A arquitetura original desta CNN contém apenas uma camada densa (a camada de saída), o que mostra que a parte convolucional foi concebida para realizar

Tabela 13 – Resultados das arquiteturas baseadas na CNN EfficientNet-B2.

Arquitetura		acc	kappa	ppv	npv	sensib	espec	AUC
SFT	efn_2	88,6 ±0,7	76,5 ±1,5	89,6 ±1,1	87,3 ±1,2	91,1 ±1,1	85,1 ±1,9	95,1 ±0,5
	efn_64	88,0 ±0,8	75,1 ±1,7	89,2 ±1,5	86,2 ±0,6	90,3 ±0,7	84,6 ±2,6	94,9 ±0,4
	efn_256	88,3 ±0,7	75,9 ±1,6	89,7 ±2,0	86,5 ±1,9	90,4 ±1,8	85,4 ±3,5	95,2 ±0,5
	efn_512	88,6 ±0,7	76,5 ±1,6	89,6 ±1,6	87,4 ±0,7	91,2 ±0,9	85,1 ±2,8	95,1 ±0,4
	efn_1024	88,5 ±0,5	76,4 ±1,2	90,3 ±1,5	86,1 ±1,2	90,0 ±1,3	86,4 ±2,5	95,0 ±0,3
DFT	efn_2	89,5 ±1,1	78,4 ±2,4	90,6 ±1,7	88,0 ±1,5	91,6 ±1,2	86,6 ±2,6	96,1 ±0,9
	efn_64	89,6 ±0,2	78,6 ±0,4	91,3 ±0,7	87,2 ±0,6	90,8 ±0,6	87,9 ±1,2	96,1 ±0,7
	efn_256	89,2 ±1,1	77,7 ±2,2	91,0 ±1,3	86,6 ±1,5	90,3 ±1,2	87,5 ±2,0	95,6 ±0,6
	efn_512	89,2 ±0,7	77,7 ±1,5	90,2 ±1,2	87,8 ±0,8	91,4 ±0,7	86,1 ±1,9	95,6 ±0,6
	efn_1024	89,3 ±0,8	78,0 ±1,8	90,7 ±1,3	87,3 ±0,3	90,9 ±0,2	87,0 ±2,1	95,3 ±0,4

– Valores em termos percentuais medidos nos conjuntos de testes.

– Os melhores resultados estão destacados em azul e os hachurados indicam que houve redução após DFT.

quase todo o processamento e já entregar uma tarefa mais simples para as camadas finais.

A arquitetura que melhor performou em termos gerais foi a icp_1024 tendo alcançado 90,7% de acurácia, 80,8% de kappa, 92,9% de sensibilidade, especificidade de 87,7% e AUC 96,6%. O estágio DFT impactou positivamente em todos os resultados obtidos com as arquiteturas baseadas na Inception-V3. A métrica mais afetada foi o kappa que aumentou 5,9 pp na arquitetura icp_1024.

Tabela 14 – Resultados das arquiteturas baseadas na CNN Inception-V3.

Arquitetura		acc	kappa	ppv	npv	sensib	espec	AUC
SFT	icp_2	87,4 ±1,6	74,0 ±3,1	88,6 ±0,4	85,7 ±3,5	89,9 ±2,9	83,9 ±0,7	94,9 ±1,0
	icp_64	88,2 ±1,5	75,6 ±2,9	89,3 ±1,2	86,7 ±3,1	90,6 ±2,5	84,8 ±1,9	95,3 ±1,0
	icp_256	88,1 ±2,2	75,5 ±4,4	89,3 ±1,6	86,7 ±4,2	90,6 ±3,6	84,8 ±2,6	95,4 ±1,5
	icp_512	88,0 ±1,5	75,4 ±2,9	89,6 ±0,8	85,9 ±3,3	89,9 ±2,8	85,4 ±1,5	95,3 ±1,4
	icp_1024	87,9 ±1,9	74,9 ±3,8	88,9 ±1,7	86,4 ±3,5	90,5 ±2,9	84,2 ±2,8	95,3 ±1,5
DFT	icp_2	89,9 ±1,7	79,1 ±3,6	91,0 ±2,0	88,3 ±2,1	91,7 ±1,7	87,2 ±3,2	96,4 ±0,6
	icp_64	90,1 ±1,9	79,7 ±4,0	91,1 ±2,1	88,7 ±1,9	92,1 ±1,3	87,4 ±3,2	96,6 ±0,7
	icp_256	90,2 ±1,2	79,9 ±2,6	91,3 ±1,1	88,7 ±1,5	92,1 ±1,1	87,7 ±1,6	96,6 ±0,7
	icp_512	90,2 ±1,7	79,8 ±3,5	91,3 ±1,8	88,7 ±2,0	92,0 ±1,5	87,7 ±2,7	96,6 ±0,8
	icp_1024	90,7 ±1,3	80,8 ±2,6	91,3 ±0,8	89,8 ±2,7	92,9 ±2,2	87,7 ±1,2	96,6 ±1,0

– Valores em termos percentuais medidos nos conjuntos de testes.

– Os melhores resultados estão destacados em azul e os hachurados indicam que houve redução após DFT.

A Tabela 15 apresenta os resultados alcançados com as arquiteturas baseadas na Resnet-50. O melhor desempenho geral foi obtido com a arquitetura rsn_1024 no estágio DFT. Aqui a acurácia medida foi de 91,7% e kappa de 83%, ppv e npv de 93,5% e 89,4%, sensibilidade e especificidade de 92,2% e 91,0%, por fim a AUC com 97,0%. Essas métricas foram as maiores atingidas entre todas as arquiteturas de base rsn, com exceção da sensibilidade, que na etapa SFT foi de 92,5%.

A continuidade do treinamento da ResNet no estágio DFT, de maneira geral

Tabela 15 – Resultados das arquiteturas baseadas na CNN ResNet-50.

Arquitetura		acc	kappa	ppv	npv	sensib	espec	AUC
SFT	rsn_2	89,9 ±1,3	79,2 ±2,7	90,8 ±2,2	88,9 ±2,4	92,1 ±2,2	86,9 ±3,7	96,2 ±0,5
	rsn_64	90,2 ±1,8	79,7 ±3,7	90,9 ±1,4	89,2 ±3,4	92,4 ±2,8	87,1 ±2,1	96,3 ±0,8
	rsn_256	89,9 ±1,1	79,2 ±2,1	90,8 ±1,8	88,9 ±3,3	92,1 ±3,1	86,8 ±3,2	96,4 ±0,7
	rsn_512	90,0 ±1,1	79,4 ±2,3	91,1 ±1,2	88,5 ±1,9	91,8 ±1,5	87,4 ±1,8	96,3 ±0,7
	rsn_1024	90,1 ±1,7	79,6 ±3,5	90,8 ±1,6	89,3 ±3,2	92,5 ±2,6	86,9 ±2,5	96,3 ±0,8
DFT	rsn_2	90,2 ±1,0	79,9 ±2,3	92,0 ±2,4	87,9 ±1,6	91,3 ±1,6	88,8 ±3,8	96,6 ±0,4
	rsn_64	91,5 ±0,9	82,6 ±1,8	93,2 ±1,6	89,3 ±2,0	92,2 ±1,8	90,5 ±2,5	97,0 ±0,5
	rsn_256	91,1 ±0,7	81,7 ±1,5	92,6 ±1,8	89,1 ±1,7	92,1 ±1,7	89,7 ±2,8	97,0 ±0,6
	rsn_512	91,5 ±0,6	82,5 ±1,2	93,2 ±0,9	89,2 ±1,9	92,1 ±1,7	90,5 ±1,5	97,0 ±0,4
	rsn_1024	91,7 ±0,9	83,0 ±1,7	93,5 ±0,8	89,4 ±2,1	92,2 ±1,8	91,0 ±1,2	97,0 ±0,4

– Valores em termos percentuais medidos nos conjuntos de testes.

– Os melhores resultados estão destacados em azul e os hachurados indicam que houve redução após DFT.

promoveu o desempenho das arquiteturas. Destacamos o kappa e a especificidade da rsn_1024 que aumentaram em 3,4 pp e 4,1 pp. Por outro lado, em alguns casos específicos, a DFT da ResNet teve efeitos negativos, de maneira similar ao observado com a DenseNet, também nas métricas de npv e sensibilidade. A maior redução foi observada na arquitetura mais simples (rsn_2) que apresentou npv com 1,0 pp a menos em relação ao que tinha obtido no estágio SFT.

A Tabela 16 apresenta os resultados observados nas arquiteturas baseadas na VGG-19. A arquitetura que se destacou no desempenho foi a vgg_256, obtendo na etapa DFT 91,1% de acurácia, 81,6% de kappa, sensibilidade de 92,1% e especificidade de 89,5%, aqui a AUC foi de 96,7%. Comparando os resultados obtidos nos estágios de treinamento, verificamos que a continuidade do treino teve efeitos positivos em todas as arquiteturas e métricas, com destaque ao kappa da vgg_2 que saiu de 65,0% na SFT para 79,7% na DFT.

Tabela 16 – Resultados das arquiteturas baseadas na CNN VGG-19.

Arquitetura		acc	kappa	ppv	npv	sensib	espec	AUC
SFT	vgg_2	83,1 ±1,5	65,0 ±3,4	84,1 ±2,6	81,8 ±0,6	87,8 ±0,8	76,5 ±4,5	91,1 ±1,3
	vgg_64	85,0 ±2,2	68,9 ±4,6	85,4 ±2,2	84,5 ±2,5	89,7 ±1,7	78,5 ±3,4	92,7 ±1,4
	vgg_256	85,1 ±1,6	69,3 ±3,3	86,5 ±1,9	83,3 ±3,0	88,4 ±2,7	80,5 ±3,4	93,0 ±1,2
	vgg_512	86,1 ±1,2	71,3 ±2,7	87,3 ±2,5	84,4 ±1,2	89,2 ±1,3	81,8 ±4,1	93,1 ±0,9
	vgg_1024	86,3 ±0,8	71,7 ±1,6	87,8 ±1,0	84,2 ±1,7	88,9 ±1,6	82,7 ±1,8	93,3 ±0,9
DFT	vgg_2	90,1 ±1,0	79,7 ±2,3	91,6 ±2,6	88,3 ±1,9	91,6 ±1,8	88,1 ±4,2	96,1 ±1,0
	vgg_64	90,5 ±1,2	80,3 ±2,6	91,7 ±2,6	88,9 ±2,1	92,1 ±1,8	88,2 ±4,3	96,6 ±1,0
	vgg_256	91,1 ±1,9	81,6 ±4,0	92,6 ±3,1	89,0 ±0,6	92,1 ±0,3	89,5 ±4,7	96,7 ±1,6
	vgg_512	90,9 ±1,2	81,2 ±2,4	92,5 ±1,5	88,7 ±1,4	91,8 ±1,1	89,5 ±2,2	96,4 ±0,8
	vgg_1024	90,4 ±1,2	80,2 ±2,5	91,4 ±1,4	89,0 ±1,4	92,2 ±1,1	87,9 ±2,1	96,2 ±1,2

– Valores em termos percentuais medidos nos conjuntos de testes.

– Os melhores resultados estão destacados em azul e os hachurados indicam que houve redução após DFT.

Observamos nos resultados individuais que a DFT, promoveu melhoria nas

métricas com algumas exceções. Para melhor visualizar esses efeitos, calculamos a razão do aumento (ou redução) em relação ao que já se tinha alcançado com a mesma arquitetura na etapa SFT $[(DFT - SFT)/SFT]$. Fizemos isso separadamente para cada métrica e calculamos a média em cada CNN-base. Dessa forma construímos a Tabela 17, que sumariza os resultados dessa análise.

Tabela 17 – Aumento médio das métricas com o treinamento da etapa DFT em relação às obtidas na SFT.

CNN-base	acc	kappa	ppv	npv	sensib	espec	AUC
DenseNet	+0,4	+1,0	+0,9	-0,3	-0,3	+1,5	+0,3
EfficientNet	+1,1	+2,6	+1,2	+0,8	+0,5	+2,0	+0,7
Inception	+2,6	+6,3	+2,3	+3,0	+2,0	+3,5	+1,4
ResNet	+1,3	+3,2	+2,2	+0,1	-0,2	+3,6	+0,6
VGG-19	+6,4	+16,6	+6,7	+6,1	+3,6	+10,9	+4,1

– Os valores estão em termos percentuais.

– Os maiores crescimentos estão destacados em azul;

– Os hachurados indicam que, na média, houve redução após DFT.

Fica perceptível que os benefícios da DFT se manifestaram de maneira bem mais expressiva na VGG-19 que nas outras CNNs-base avaliadas, com destaque para o kappa que aumentou em média 16,6% em relação à SFT. A topologia da VGG-19 é simples (tipo sequencial), e comparada com as outras CNNs avaliadas, originalmente possuía muito mais parâmetros na parte densa (cerca de 86% dos seus parâmetros, conforme Tabela 9). Acreditamos que ao substituir a parte densa original por uma bem menor com apenas uma ou duas camadas e no máximo 1.026 neurônios, essa nova parte densa não foi capaz de realizar bem as transformações necessárias nos atributos recebidos da parte convolucional.

Assim, ao descongelar o treinamento também na parte convolucional (na DFT), o ajuste geral dos pesos da rede proporcionou uma melhor sintonia entre as partes convolucional e densa das arquiteturas, desencadeando uma adaptação mais profunda do aprendizado aos novos dados.

5.2 Seleção de modelos e formação de *ensembles*

Na seção anterior, mostramos os resultados de cada classificador individualmente aferidos no conjunto de testes. Para escolher os membros que formarão os comitês, usamos como métrica de comparação de modelos, a AUC média no conjunto de validação. Dessa forma podemos ordenar os modelos de forma objetiva, por meio de uma métrica que além de levar em conta todos os quadrantes da matriz de confusão, verifica o quão bem o modelo separa as classes positiva e negativa.

A imagem da Figura 28 apresenta a métrica AUC obtida por cada modelo medida nos conjuntos de validação. Pelo gráfico conseguimos visualmente identificar que, assim como aconteceu no conjunto de testes, a OST no conjunto de validação também teve desempenho bem menor em relação aos obtidos com a estratégia de duas etapas. No mesmo caminho, a DFT também promoveu melhorias sobre a SFT no conjunto de validação.

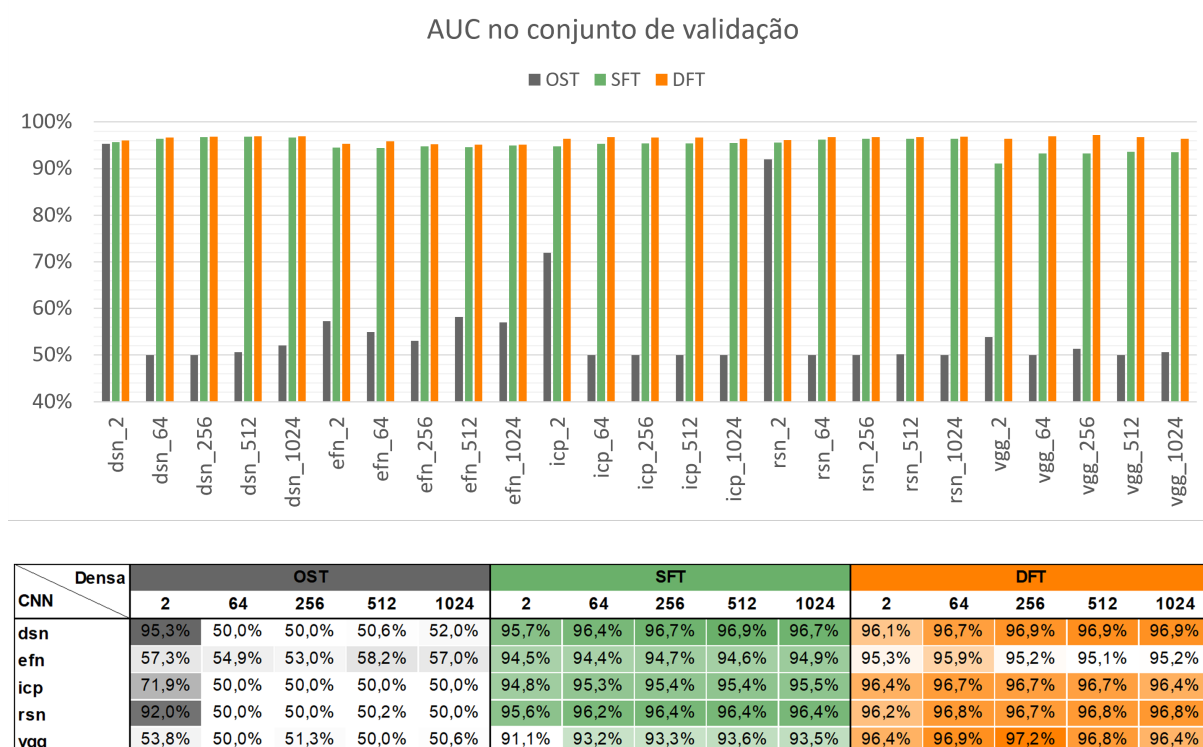


Figura 28 – Gráfico e tabela de dados da AUC de validação nas etapas de treinamento.

Tendo ranqueado os modelos com base na AUC aferida no conjunto de validação, observamos os critérios de seleção de membros para formação de comitês discutidos na Seção 4.4, formamos os grupos dispostos na Tabela 18. Os classificadores de cada grupo da tabela estão em ordem decrescente de AUC, portanto, os comitês de três membros são formados apenas pelos três primeiros classificadores de cada grupo correspondente.

Entre os critérios de formação dos *ensembles* há um que determina a etapa de treinamento: OST, SFT, DFT ou em qualquer etapa. Em decorrência dos resultados individuais, no grupo 5Ad, que não restringe o estágio de treino, apenas classificadores do DFT foram selecionados, tornando-se igual ao grupo 5Dd. Consequentemente, o mesmo acontece com 3Ad e 3Dd que contém os três primeiros classificadores de cada grupo correspondente. Outros pares de grupos semelhantes são: 3Od e 3Or; 3Dr e 3Ar. Portanto, para evitar redundância no texto, não reportaremos os resultados dos grupos 5Ad, 3Ad, 3Or e 3Ar.

Tabela 18 – *Ensembles* formados e seus membros classificadores.

Grupo	Classificadores				
	I	II	III	IV	V
5Od	dsn_2	rsn_2	icp_2	efn_512	vgg_2
5Or	dsn_2	rsn_2	icp_2	efn_512	efn_2
5Sd	dsn_512	rsn_512	icp_1024	efn_1024	vgg_512
5Sr	dsn_512	dsn_256	dsn_1024	rsn_512	rsn_256
5Dd	vgg_256	dsn_512	rsn_1024	icp_64	efn_64
5Dr	vgg_256	dsn_512	vgg_64	dsn_1024	dsn_256
5Ad	vgg_256	dsn_512	rsn_1024	icp_64	efn_64
5Ar	vgg_256	dsn_512	vgg_64	dsn_1024	dsn_512

– Cores: cinza indica etapa OST; verde - SFT; laranja - DFT. Quando aplicada ao nome do grupo, indica que todos os membros são obrigatoriamente da mesma etapa.

– Os comitês de três membros são formados pelos três primeiros classificadores de cada grupo correspondente.

5.3 Resultados com *ensembles*

Nesta seção, mostramos os resultados obtidos pelos *ensembles*. Cada grupo formado produz um conjunto de classificações que são combinadas por uma função de apuração, conforme descrito na Seção 4.4. Ao todo temos 12 grupos de classificadores distintos e 9 funções de apuração diferentes, o que resulta em 108 comitês. Para comparar o desempenho entre eles mantivemos a mesma estratégia utilizada na seleção de membros: baseado na AUC no conjunto de validação. A Figura 29 apresenta então a AUC média nos conjuntos de validação que cada comitê atingiu.

Grupo	Funções de apuração/cominação								
	P ₀ W ₀	P ₀ W ₁	P ₀ W ₂	P ₁ W ₀	P ₁ W ₁	P ₁ W ₂	P ₂ W ₀	P ₂ W ₁	P ₂ W ₂
5Od	90,53%	90,98%	90,98%	92,00%	92,04%	92,03%	95,19%	95,22%	95,20%
5Or	90,59%	91,01%	91,01%	92,00%	92,03%	92,04%	95,20%	95,23%	95,20%
3Od	90,74%	90,98%	90,98%	92,05%	92,10%	92,09%	95,13%	95,17%	95,17%
3Sr	91,71%	91,71%	91,71%	91,69%	91,69%	91,69%	96,81%	96,81%	96,81%
5Sr	94,95%	94,99%	94,98%	94,97%	94,96%	94,97%	97,29%	97,29%	97,29%
3Sd	95,38%	95,48%	95,48%	95,45%	95,46%	95,46%	97,43%	97,43%	97,43%
5Dr	95,38%	95,40%	95,40%	95,46%	95,47%	95,47%	97,56%	97,56%	97,56%
3Dr	95,41%	95,38%	95,38%	95,52%	95,52%	95,52%	97,59%	97,60%	97,60%
5Ar	95,85%	95,88%	95,90%	95,94%	95,94%	95,94%	97,60%	97,60%	97,61%
3Dd	96,09%	96,11%	96,11%	96,23%	96,22%	96,22%	97,77%	97,77%	97,77%
5Sd	96,24%	96,37%	96,37%	96,33%	96,35%	96,35%	97,46%	97,45%	97,46%
5Dd	97,07%	97,12%	97,13%	97,13%	97,13%	97,13%	97,82%	97,82%	97,82%

Menor auc

Maior auc

Figura 29 – Métrica AUC obtidas pelos comitês no conjunto de validação.

A Figura 29 constitui uma ferramenta interessante para construir algumas análises de forma visual. De início vemos que os comitês formados por classificadores no estágio OST obtiveram os menores desempenhos. Essa observação está consoante ao baixo rendimento dos seus membros, visto que cada classificador individual deve ser bem treinado e ter um bom desempenho para que o *ensemble* como um todo seja eficaz.

Em outra análise, comparando o desempenho entre as funções de apuração, percebemos uma tendência de crescimento da AUC conforme se usa as probabilidades de saída (*softmax*) dos classificadores: $P_0 < P_1 < P_2$. O mesmo efeito não foi observado quanto a aplicação de pesos, (W_0 , W_1 e W_2), estes impactaram com menor força sobre a AUC, em muitos casos sendo percebidos somente em casas decimais mais precisas.

Outra observação é que, com exceção dos grupos de etapa OST, observa-se que a regra de obrigatoriedade de CNNs de bases distintas (grupos terminados em *d*) apresentaram maior AUC que seu correspondente que permite bases repetidas (grupos terminados em *r*). Ex: 3Sr < 3Sd, 5Sr < 5Sd, 5Dr < 5Dd. O que aponta para os benefícios advindos da variedade dos membros votantes.

Os comitês do grupo 5Dd, com as funções de apuração que consideram ambas as probabilidades do *softmax* (P_2) alcançaram a maior AUC de validação (97,82%). Então para selecionar o melhor comitê, analisamos adicionalmente a acurácia e o kappa no mesmo conjunto, pois também são calculados com toda a matriz de confusão. Em decorrência dessa análise, foi decido pela função P_2W_2 , pois apresentou menores desvios padrão e maior acc.

A Tabela 19 apresenta os resultados obtidos com estes três *ensembles* no conjunto de testes. Similar ao observado no conjunto de validação, percebemos um equilíbrio no desempenho deles. Observamos melhorias de todas as métricas em comparação com os classificadores individuais que compõem o comitê. Além disso, na maioria dos casos houve redução do desvio padrão, o que indica que os *ensembles* produzem resultados mais estáveis para os diferentes conjuntos de imagens de teste.

Tabela 19 – Resultados dos comitês do grupo 5Dd com as funções de apuração tipo P_2 no conjunto de testes.

	Comitê	acc	kappa	ppv	npv	sensib	espec	AUC
5Dd	P_2W_0	92,8 ± 0,7	85,1 ± 1,5	94,1 ± 0,9	91,0 ± 0,9	93,5 ± 0,8	91,8 ± 1,3	98,1 ± 0,3
	P_2W_1	92,8 ± 0,7	85,1 ± 1,5	94,1 ± 0,9	91,0 ± 0,9	93,5 ± 0,8	91,8 ± 1,3	98,1 ± 0,3
	P_2W_2	92,8 ± 0,8	85,2 ± 1,7	94,0 ± 1,0	91,2 ± 1,0	93,6 ± 0,8	91,7 ± 1,4	98,1 ± 0,3

– Valores em termos percentuais medidos nos conjuntos de testes.

Construímos o gráfico da Figura 30 para permitir comparar o desempenho geral do *ensemble* 5Dd combinado via P_2W_2 com o de seus membros individualmente. Percebe-se que a linha do *ensemble* no radar circunda as demais devido a sua performance superar

a de seus membros individualmente em todas as métricas.

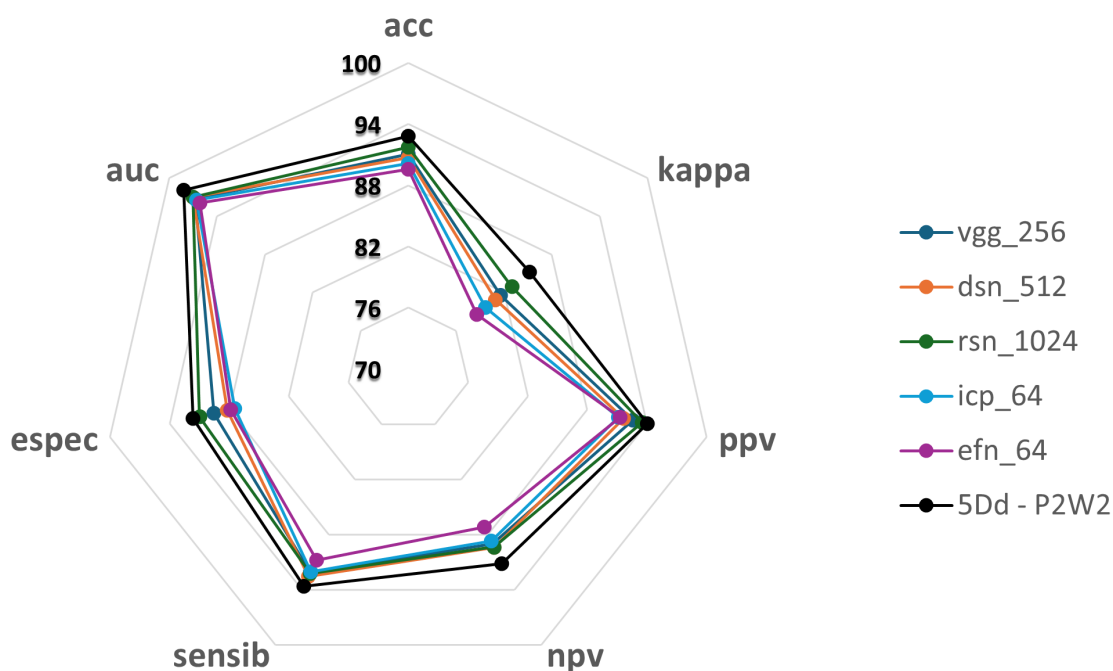


Figura 30 – Comparação de métricas do *ensemble* 5Dd - P₂W₂ e de seus membros individuais no conjunto de testes.

Além da comparação direta, confrontamos as métricas do comitê às dos classificadores individuais por meio de testes estatísticos. Aferimos a probabilidade associada ao teste t de Student, unicaudal para amostras pareadas com nível de significância de 0,05. Os resultados desse teste apontaram que as métricas médias do *ensemble* foram significativamente maiores em 30 comparações. Já nestas cinco: ppv e especificidade da vgg_256 e rs_n_1024, e na sensibilidade da dsn_512, o teste apontou que não há evidência suficiente para concluir que há diferença nas médias.

A Figura 31 apresenta uma matriz de confusão contendo algumas predições do *ensemble* 5Dd - P₂W₂ nos conjuntos de testes. Para verificar o desempenho do *ensemble* específico por corante, anotamos a quantidade de VP, VN, FP e FN separadamente para cada coloração. A acurácia considerando apenas as imagens tratadas com H&E foi de 94,3%, em imagens com PAS, 91,0% e nas imagens coradas com PMJ, 91,3%.

No sentido de avançar rumo a explicabilidade dos modelos, executamos o método de XAI GRAD-CAM sobre as imagens de teste. Dessa forma, além de dispor da classificação final predita pelo modelo, podemos ver as regiões que mais influenciaram na classificação por ele fornecida.

A Figura 32 apresenta os mapas de atenção gerados com o GRAD-CAM realizado em cada CNN votante do *ensemble* 5Dd - P₂W₂, sobre duas imagens da nossa base. Na

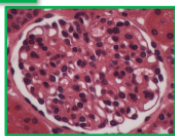
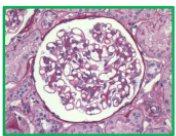
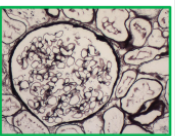
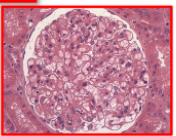
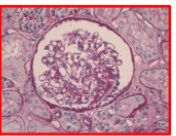
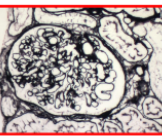
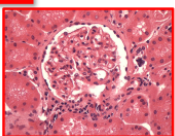
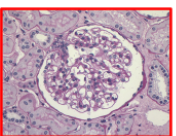
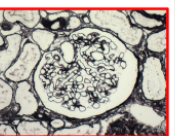
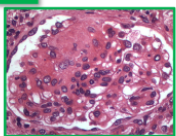
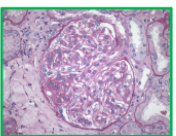
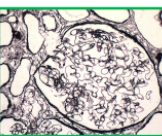
		CLASSIFICAÇÃO REAL							
		Saudável			GS				
CLASSIFICAÇÃO 5Dd – P ₂ W ₂	Saudável	VN				FN			
			H&E 494	PAS 202	PMJ 128		H&E 35	PAS 28	PMJ 17
	GS	FP				VP			
			H&E 32	PAS 30	PMJ 13		H&E 611	PAS 382	PMJ 186

Figura 31 – Amostra de imagens classificadas com o *ensemble* 5Dd - P₂W₂ no conjunto de testes. Indicamos a quantidade de cada tipo de classificação (VP, VN, FP, FN) separadamente por corante.

primeira observa-se que todas as redes deram muita atenção a região do glomérulo e pouca relevância para as regiões externas. Isso mostra que as redes selecionaram os locais de atenção de forma correta e em consonância ao que faz o especialista.

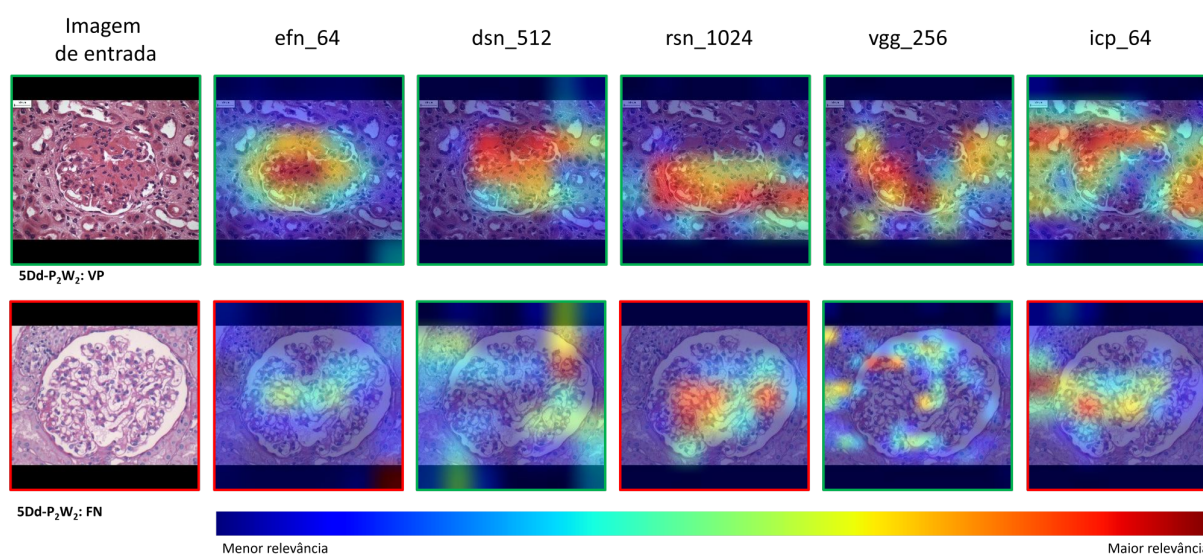


Figura 32 – Mapas de atenção produzidos com o GradCam das CNNs votantes no *ensemble* 5Dd - P₂W₂. Os mapas foram sobrepostos à imagem original para facilitar a visualização de áreas correspondentes. Os tons de cor vão do azul ao vermelho conforme a relevância da região para a classificação. A borda verde ou vermelha na imagem indica a classificação correta ou incorreta, respectivamente.

Outro aspecto é que mesmo dando relevância ao glomérulo, os mapas de atenção são diferentes entre si, ressaltando que cada rede pondera as características de formas distintas. Uma vez que o método se baseia não nas observações isoladas de uma única

CNN, mas sim via comitê, esse fato faz com que haja mais riqueza de características consideradas para classificação.

Já na segunda imagem da Figura 32, os mapas de atenção mostram que as regiões consideradas de maior importância para a classificação estão difusas, alguns pontos de atenção estão fora do glomérulo ou só focam em uma porção muito pequena do mesmo.

Também investigamos a possibilidade de *overfitting* das CNNs que compõem o *ensemble*. De acordo com Shorten e Khoshgoftaar (2019), uma maneira de descobrir o *overfitting* é plotar a acurácia nos conjuntos de treino e validação em cada época durante o treinamento das redes. A acurácia do conjunto de validação deve continuar a aumentar junto com a do conjunto de teste.

A Figura 33 apresenta gráficos da *loss* e acurácia das cinco CNNs que compõem o *ensemble*, aferidas nesses dois conjuntos. Os modelos utilizados foram treinados até a época de menor *loss* de validação. Podemos verificar que do início do treinamento até essa época as medições nos conjuntos de validação e treino seguem juntas na mesma tendência de crescimento. Embora apresentem alguns picos, existe uma relação entre as duas curvas, não havendo pontos de inflexão onde a acurácia no treino continua a aumentar e a acurácia da validação começa a diminuir, afastando a característica que marca o *overfitting*.

Os resultados obtidos pelo *ensemble* 5Dd - P₂W₂ alcançaram métricas em patamares compatíveis com outras propostas encontradas em trabalhos relacionados focados na classificação de glomérulos saudáveis ou com esclerose. Com destaque para a maior variedade de corantes presentes no dataset bem como a quantidade de imagens da classe positiva e relativo balanceamento de classes. A Tabela 20 apresenta dados que permitem essa comparação.

Tabela 20 – Comparação do método proposto com outros trabalhos relacionados.

Trabalho	Dataset	Corantes	Resultados
Araújo et al. (2017)	747 56%	PAS, H&E	acc: 84,8% (H&E); 81,3% (PAS)
Marsh et al. (2018)	3.867 22%	H&E	ppv: 89,3%; sensib: 86,5%
Kannan et al. (2019)	745 18%	Tricrômico	acc: 92,7%, kappa: 86,8%
Pesce et al. (2022)	2.772 15%	PAS	ppv: 98,4%; sensib: 94,3%
5Dd - P ₂ W ₂ (proposto)	2.158 58%	H&E, PAS, PMJ	acc: 92,8%; kappa: 85,2; ppv: 94,0%; sensib: 93,6%

– No dataset indicamos o total de imagens de glomérulos e a porcentagem da classe positiva.

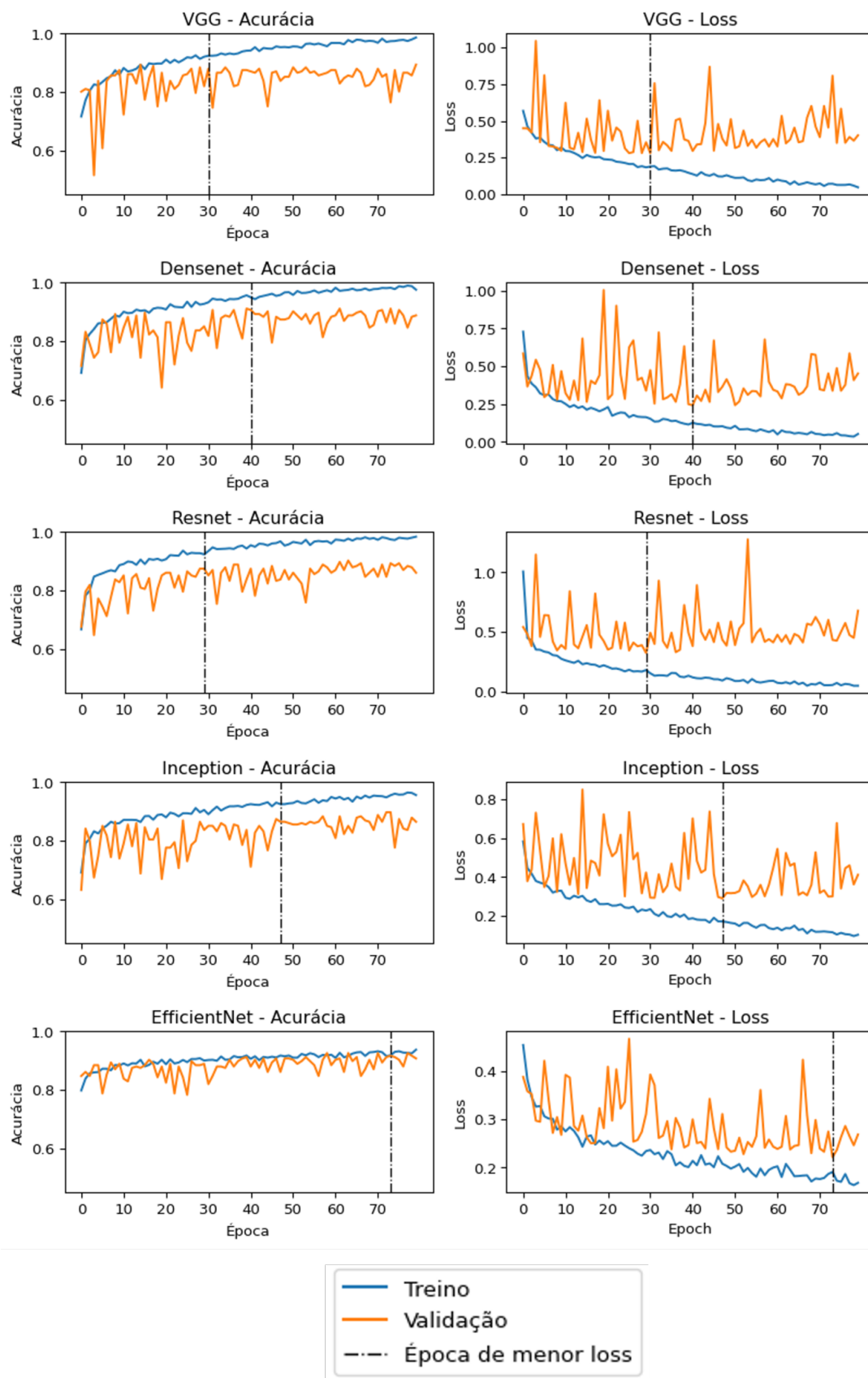


Figura 33 – *Loss* e acurácia das CNNs do *ensemble* 5Dd - P_2W_2 em cada época nos conjuntos de validação e treino.

5.4 Considerações Finais

Nesse capítulo, apresentamos e analisamos os resultados obtidos durante os experimentos realizados no desenvolvimento desta proposta. Adaptamos e avaliamos o desempenho de 25 arquiteturas de CNNs. Realizamos os treinamentos em estágio único e em dois estágios. Combinamos os classificadores em vários comitês com diferentes mecanismos de apuração.

Dessa forma, chegamos à proposta do *ensemble* 5Dd, formado por cinco arquiteturas adaptadas das CNNs DenseNet, EfficientNet, Inception, ResNet e VGG, considerando o mecanismo de apuração probabilístico e ponderado (P_2W_2). O método se destacou nos resultados alcançados, superando as demais alternativas avaliadas não somente nas métricas aferidas, mas também em estabilidade nos diferentes conjuntos de testes. No próximo capítulo, serão apresentadas as conclusões deste trabalho.

6 Conclusão

O presente trabalho está inserido no contexto de sistemas de diagnóstico auxiliado por computador. Desenvolvemos um método que classifica imagens histológicas de biópsia de glomérulos em saudáveis ou com glomerulosclerose. O método desenvolvido foi avaliado em amostras preparadas com corantes químicos diferentes (H&E, PAS, E Pratas Metenamina e Jones) e não necessita de ajuste de parâmetros para fornecer a classificação.

Aplicamos técnicas de transferência de aprendizado sobre redes neurais convolucionais já estabelecidas no meio acadêmico-científico. Aproveitamos a alta capacidade de generalização e extração de características de cinco CNNs com topologias variadas, Densenet, EfficientNet, Inception, Resnet e VGG e adaptamos suas arquiteturas para realizar a tarefa específica de classificação de biópsias renais para o diagnóstico de glomeruloesclerose.

Realizamos aumento de dados sobre o dataset de treinamento dos modelos gerando mais imagens artificialmente. Implementamos transformações que refletem situações eventuais do processo de captura de imagem e da preparação e análise laboratorial. Ruídos decorrentes de variações da intensidade de iluminação do microscópio e da concentração ou tonalidade corantes químicos, foram representadas pela adição e/ou redução de brilho e contraste nas imagens. Também foram aplicadas transformações espaciais cautelosamente definidas para não prejudicarem as características essenciais do tecido (*label preserving transformations*).

Após avaliar as estratégias de treinamento em etapa única (OST) e em dois estágios (SFT e DFT), concluímos que é mais eficiente o treinamento em duas etapas, considerando que a arquitetura adaptada inicia parcialmente pré-treinada na ImageNet (somente os pesos convolucionais são importados). A avaliação de desempenho dos modelos foi realizada separadamente em cada etapa. A SFT já apresenta resultados maiores que a OST e a continuidade de treino com a DFT ainda melhora as métricas aferidas de maneira geral.

Em todas as estratégias mencionadas, cuidamos para não utilizar modelos hiperajustados ao conjunto de treino (*overfitting*) e de baixa capacidade de generalização. Para isso, a cada época foi feito o acompanhamento da *loss* e acurácia no conjuntos de treino e validação. Os modelos utilizados são aqueles da época de menor *loss* de validação.

Foram formados comitês agrupando os melhores modelos com base na AUC de validação. Os critérios de agrupando definidos promoveram a formação de comitês de

diferentes características no que diz respeito a quantidade de membros, à variedade de arquiteturas, e ao estágio de treinamento. Quanto a função de combinação das predições, além do voto majoritário, implementamos mais oito formatos, levando em consideração a saída probabilística das CNNs, bem como mecanismos de votação ponderada.

Analizando os resultados obtidos com os comitês, quanto a quantidade de membros, em geral os comitês com 5 votantes tiveram desempenho maior em relação aos de 3 votantes. No que se refere ao estágio de treinamento, os comitês seguiram a mesma tendência observada nos classificadores individuais: OST < SFT < DFT, aqueles contendo membros em diferentes estágios performaram de forma intermediária. Por fim os grupos que continham mais variedades de CNNs-base performaram melhor em relação ao que continha arquiteturas de mesma base. Em suma, a melhor performance dos modelos votantes, a quantidade e variedade de membros promovem comitês mais eficazes.

No que se refere aos mecanismos de combinação de predições nos *ensembles*, verificamos que a forma de apuração que considerando a distribuição de probabilidade da saída das CNNs promove uma melhor separação das classes do métodos categóricos mais tradicionais, como o voto majoritário por exemplo. Por outro lado, a ponderação de voto usando as métricas de cada classificador como fator multiplicativo, apresentou para este caso efeito difuso e de menor impacto.

Os resultados alcançados com o *ensemble* proposto foram medidos com 7 métricas (acurácia, kappa, ppv, npv, sensibilidade, especificidade e AUC). A performance obtida pelo comitês de classificação foi maior que a alcançada por qualquer um dos seus respectivos classificadores membros individualmente.

O método proposto apresenta resultados que o coloca em patamares semelhantes aos de trabalhos relacionados da literatura estudada, e ainda se destaca nos seguintes aspectos: Variedade de corantes na base de dados; Quantidade de arquiteturas utilizadas; análise de diferentes estratégias de treinamento; variedade de comitês formados e mecanismos de combinação de predições.

6.1 Trabalhos Futuros

Embora os resultados obtidos para a detecção de glomeruloesclerose sejam satisfatórios, ainda há espaço para melhorias no método proposto. A seguir, destacam-se algumas sugestões:

1. Explorar o espaço de hiperparâmetros de treinamento e das próprias CNNs para buscar configurações que incrementem o desempenho dos métodos.
2. Elaborar novas arquiteturas de redes neurais convolucionais desenhadas especificamente para trabalhar com histopatologia renal.

3. Desenvolver e avaliar classificadores que façam uso de outras características mais tradicionais como descritores de textura da imagem, por exemplo.
4. Combinar classificadores por técnicas diferentes do comitê, por exemplo multilevel.
5. Avançar no sentido da aplicação de técnicas de explicabilidade em IA, ao método proposto.

Referências

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **Software available from tensorflow.org**, 2015. Citado na página 63.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37–66, Jan 1991. Citado na página 40.

ALVES, E.; FILHO, J. B. S.; KRITSKI, A. L. An ensemble approach for supporting the respiratory isolation of presumed tuberculosis inpatients. **Neurocomputing**, Elsevier, v. 331, p. 289–300, 2019. Citado na página 25.

ARAÚJO, I. C.; SCHNITMAN, L.; DUARTE, A. A.; SANTOS, W. dos. Automated detection of segmental glomerulosclerosis in kidney histopathology. In: **XIII Brazilian Congress on Computational Intelligence**. [S.l.: s.n.], 2017. p. 12. Citado 3 vezes nas páginas 42, 43 e 75.

BARROS, G. O.; NAVARRO, B.; DUARTE, A.; DOS-SANTOS, W. L. Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys. **Scientific reports**, Nature Publishing Group, v. 7, n. 1, p. 1–8, 2017. Citado 3 vezes nas páginas 40, 41 e 43.

BARROS, G. O.; WANDERLEY, D. C.; REBOUÇAS, L. O.; SANTOS, W. L. dos; DUARTE, A. A.; VIDAL, F. de B. Podnet: Ensemble-based classification of podocytopathy on kidney glomerular images. In: **VISIGRAPP (5: VISAPP)**. [S.l.: s.n.], 2022. p. 405–412. Citado 2 vezes nas páginas 41 e 43.

BEVILACQUA, V.; PIETROLEONARDO, N.; TRIGGIANI, V.; BRUNETTI, A.; Di Palma, A. M.; ROSSINI, M.; GESUALDO, L. An innovative neural network framework to classify blood vessels and tubules based on haralick features evaluated in histological images of kidney biopsy. **Neurocomputing**, v. 228, p. 143–153, 2017. ISSN 0925-2312. Advanced Intelligent Computing: Theory and Applications. Citado na página 40.

BUENO, G.; FERNANDEZ-CARROBLES, M. M.; GONZALEZ-LOPEZ, L.; DENIZ, O. Glomerulosclerosis identification in whole slide images using semantic segmentation. **Computer methods and programs in biomedicine**, Elsevier, v. 184, p. 105273, 2020. Citado 3 vezes nas páginas 16, 22 e 40.

CHAGAS, P.; SOUZA, L.; ARAÚJO, I.; ALDEMAN, N.; DUARTE, A.; ANGELO, M.; DOS-SANTOS, W. L.; OLIVEIRA, L. Classification of glomerular hypercellularity using convolutional features and support vector machine. **Artificial intelligence in medicine**, Elsevier, v. 103, p. 101808, 2020. Citado 2 vezes nas páginas 41 e 43.

CHOLLET, F. et al. **Keras**. 2015. <<https://keras.io>>. Citado na página 63.

CLARO, M. d. L. **Classificação de Leucemias Utilizando Aumento de Dados, Transferência de Aprendizado e Combinação de CNNs**. Tese (Doutorado) — Universidade Federal do Piauí, 2022. Citado na página 25.

CLARO, M. d. L. et al. Classificação de leucemias utilizando aumento de dados, transferência de aprendizado e combinação de cnns. Universidade Federal do Maranhão, 2022. Citado na página 49.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, p. 37–46, 1960. Citado na página 37.

COSTA, D. M. d. N.; VALENTE, L. M.; GOUVEIA, P. A. d. C.; SARINHO, F. W.; FERNANDES, G. V.; CAVALCANTE, M. A. G. d. M.; OLIVEIRA, C. B. L. d.; VASCONCELOS, C. d. A. J. d.; SARINHO, E. S. C. Comparative analysis of primary and secondary glomerulopathies in the northeast of Brazil: data from the Pernambuco registry of glomerulopathies-repeg. **Brazilian Journal of Nephrology**, SciELO Brasil, v. 39, n. 1, p. 29–35, 2017. Citado na página 20.

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 25.

DAVITA INC. **kidney biopsy**. 2023. Disponível em: <<https://www.davita.com/education/kidney-vocabulary/kidney-biopsy>>. Acesso em: 26 de fevereiro de 2023. Citado na página 21.

DAVITA INC. **What Is Glomerulosclerosis?** 2023. Disponível em: <<http://www.davita.com/education/kidney-disease/symptoms/what-is-glomerulosclerosis>>. Acesso em: 30 de janeiro de 2023. Citado 2 vezes nas páginas 16 e 22.

DHAUN, N.; BELLAMY, C.; CATTRAN, D.; KLUTH, D. Utility of renal biopsy in the clinical management of renal disease: hematuria should not be missed reply. **Kidney International**, Elsevier Inc., v. 86, n. 6, p. 1269–1269, dez. 2014. ISSN 0085-2538. Citado na página 21.

DIETTERICH, T. G. Ensemble methods in machine learning. In: **Multiple Classifier Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. p. 1–15. ISBN 978-3-540-45014-6. Citado na página 25.

DOI, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. **Computerized Medical Imaging and Graphics**, p. 198–211, 2007. Citado na página 17.

FERRON, M.; RANCANO, J. **Grande Atlas Do Corpo Humano**. [S.l.]: MANOLE, 2007. ISBN 9788520424995. Citado na página 20.

FUND, W. C. R. **Kidney Cancer Statistics**. 2019. Disponível em: <<https://www.wcrf.org/dietandcancer/cancer-trends/kidney-cancer-statistics>>. Acesso em: 27 de nov. de 2020. Citado na página 16.

GINLEY, B.; LUTNICK, B.; JEN, K.-Y.; FOGO, A. B.; JAIN, S.; ROSENBERG, A. Z.; WALAVALKAR, V.; WILDING, G. E.; TOMASZEWSKI, J. E.; YACOB, R.; ROSSI, G. M.; SARDER, P. Computational segmentation and classification of diabetic glomerulosclerosis. **Journal of the American Society of Nephrology : JASN**, 2019. Citado na página 40.

- GINLEY, B. G.; TOMASZEWSKI, J. E.; JEN, K.-Y.; FOGO, A.; JAIN, S.; SARDER, P. Computational analysis of the structural progression of human glomeruli in diabetic nephropathy. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2018: Digital Pathology**. [S.l.], 2018. v. 10581, p. 105810A. Citado 3 vezes nas páginas 17, 40 e 43.
- GONZALEZ, R.; WOODS, R. **Digital image processing**. [S.l.]: Pearson, Prentice Hall, 2008. 954 p. Citado 3 vezes nas páginas 22, 23 e 50.
- HARALICK, R. M.; SHANMUGAM, K. et al. Textural features for image classification. **IEEE Transactions on systems, man, and cybernetics**, IEEE, v. 3, n. 6, p. 610–621, 1973. Citado na página 41.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 41.
- HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.], 2015. p. 770–778. Citado 3 vezes nas páginas 29, 30 e 52.
- HECKENAUER, R.; WEBER, J.; WEMMERT, C.; FEUERHAKE, F.; HASSENFORDER, M.; MULLER, P.-A.; FORESTIER, G. Real-time detection of glomeruli in renal pathology. In: IEEE. **2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)**. [S.l.], 2020. p. 350–355. Citado na página 40.
- HUANG, G.; LIU, Z.; WEINBERGER, K. Q. Densely connected convolutional networks. **CoRR**, abs/1608.06993, 2016. Citado 3 vezes nas páginas 31, 32 e 52.
- HUO, Y.; DENG, R.; LIU, Q.; FOGO, A. B.; YANG, H. Ai applications in renal pathology. **Kidney international**, Elsevier, 2021. Citado na página 17.
- ISAAKS, E.; SRIVASTAVA, R.; (FIRM), K. **An introduction to Applied Geostatistics**. [S.l.]: Oxford University Press, 1989. ISBN 9780195050134. Citado na página 41.
- JONES, E.; OLIPHANT, T.; PETERSON, P. et al. **SciPy: Open source scientific tools for Python**. 2001. Disponível em: <<http://www.scipy.org/>>. Citado na página 62.
- KALANTARI, A.; KAMSIN, A.; SHAMSHIRBAND, S.; GANI, A.; ALINEJAD-ROKNY, H.; CHRONOPOULOS, A. T. Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. **Neurocomputing**, v. 276, p. 2–22, 2018. ISSN 0925-2312. Machine Learning and Data Mining Techniques for Medical Complex Data Analysis. Citado na página 28.
- KANNAN, S.; MORGAN, L. A.; LIANG, B.; CHEUNG, M. G.; LIN, C. Q.; MUN, D.; NADER, R. G.; BELGHASEM, M. E.; HENDERSON, J. M.; FRANCIS, J. M.; CHITALIA, V. C.; KOLACHALAMA, V. B. Segmentation of glomeruli within trichrome images using deep learning. **Kidney International Reports**, v. 4, n. 7, p. 955–962, 2019. ISSN 2468-0249. Citado 3 vezes nas páginas 42, 43 e 75.

KOLACHALAMA, V. B.; SINGH, P.; LIN, C. Q.; MUN, D.; BELGHASEM, M. E.; HENDERSON, J. M.; FRANCIS, J. M.; SALANT, D. J.; CHITALIA, V. C. Association of pathological fibrosis with renal survival using deep neural networks. **Kidney International Reports**, v. 3, n. 2, p. 464–475, 2018. ISSN 2468-0249. Citado na página 40.

KORNBLITH, S.; SHLENS, J.; LE, Q. V. Do better imagenet models transfer better? In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2019. p. 2661–2671. Citado na página 52.

KOTYK, T.; DEY, N.; ASHOUR, A. S.; BALAS-TIMAR, D.; CHAKRABORTY, S.; ASHOUR, A. S.; TAVARES, J. M. R. Measurement of glomerulus diameter and bowman's space width of renal albino rats. **Computer Methods and Programs in Biomedicine**, v. 126, p. 143 – 153, 2016. ISSN 0169-2607. Citado na página 17.

LAAK, J. Van der; LITJENS, G.; CIOMPI, F. Deep learning in histopathology: the path to the clinic. **Nature Medicine**, v. 27, n. 5, p. 775–784, 2021. Citado na página 17.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, JSTOR, p. 159–174, 1977. Citado na página 37.

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 25.

LI, S.; GONG, K.; LIU, C. H.; WANG, Y.; QIAO, F.; CHENG, X. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2021. p. 5212–5221. Citado na página 48.

MARSH, J. N.; MATLOCK, M. K.; KUDOSE, S.; LIU, T.-C.; STAPPENBECK, T. S.; GAUT, J. P.; SWAMIDASS, S. J. Deep learning global glomerulosclerosis in transplant kidney frozen sections. **IEEE transactions on medical imaging**, IEEE, v. 37, n. 12, p. 2718–2728, 2018. Citado 5 vezes nas páginas 17, 21, 42, 43 e 75.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 25.

MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. Methods for interpreting and understanding deep neural networks. **Digital signal processing**, Elsevier, v. 73, p. 1–15, 2018. Citado na página 36.

NERBASS, F. B.; LIMA, H. d. N.; THOMÉ, F. S.; NETO, O. M. V.; LUGON, J. R.; SESSO, R. Brazilian dialysis survey 2020. **Brazilian Journal of Nephrology**, SciELO Brasil, v. 44, p. 349–357, 2022. Citado na página 16.

NEWMAN, T. **What do the kidneys do?** 2019. Disponível em: <<https://www.medicalnewstoday.com/articles/305488>>. Acesso em: 03 de maio de 2021. Citado na página 20.

OLIVEIRA, L.; CHAGAS, P.; DUARTE, A.; CALUMBY, R.; SANTOS, E.; ANGELO, M.; SANTOS, W. dos. Pathospotter: Computational intelligence applied to nephropathology. In: _____. **Innovations in Nephrology: Breakthrough Technologies in Kidney**

Disease Care. Cham: Springer International Publishing, 2022. p. 253–272. ISBN 978-3-031-11570-7. Citado na página 28.

O'SHAUGHNESSY, M. M.; HOGAN, S. L.; THOMPSON, B. D.; COPPO, R.; FOGO, A. B.; JENNETTE, J. C. Glomerular disease frequencies by race, sex and region: results from the International Kidney Biopsy Survey. **Nephrology Dialysis Transplantation**, v. 33, n. 4, p. 661–669, 07 2017. ISSN 0931-0509. Disponível em: <<https://doi.org/10.1093/ndt/gfx189>>. Citado na página 16.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, 2010. Citado na página 24.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 62.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. [S.l.]: Thomson Learning, 2008. Citado na página 22.

PESCE, F.; ALBANESE, F.; MALLARDI, D.; ROSSINI, M.; PASCULLI, G.; SUAVO-BULZIS, P.; GRANATA, A.; BRUNETTI, A.; CASCARANO, G. D.; BEVILACQUA, V.; GESUALDO, L. Identification of glomerulosclerosis using ibm watson and shallow neural networks. **Journal of Nephrology**, v. 35, n. 4, p. 1235–1242, 2022. Citado 3 vezes nas páginas 42, 43 e 75.

PICCIALLI, F.; GIAMPAOLO, F.; SALVI, A.; CUOMO, S. A robust ensemble technique in forecasting workload of local healthcare departments. **Neurocomputing**, Elsevier, v. 444, p. 69–78, 2021. Citado na página 25.

RAS, G.; GERVEN, M. van; HASELAGER, P. Explanation methods in deep learning: Users, values, concerns and challenges. **Explainable and interpretable models in computer vision and machine learning**, Springer, p. 19–36, 2018. Citado na página 36.

RAS, G.; XIE, N.; GERVEN, M. V.; DORAN, D. Explainable deep learning: A field guide for the uninitiated. **Journal of Artificial Intelligence Research**, v. 73, p. 329–397, 2022. Citado na página 36.

REHEM, J. M. C.; SANTOS, W. L. C. dos; DUARTE, A. A.; OLIVEIRA, L. R. de; ANGELO, M. F. Automatic glomerulus detection in renal histological images. In: SPIE. **Medical Imaging 2021: Digital Pathology**. [S.l.], 2021. v. 11603, p. 115–125. Citado na página 40.

RIBANI, R.; MARENGONI, M. A survey of transfer learning for convolutional neural networks. In: **2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)**. [S.l.: s.n.], 2019. p. 47–57. Citado na página 24.

RISDON, R. A.; TURNER, D. R. **Atlas of renal pathology**. [S.l.]: Springer Science & Business Media, 2012. v. 2. Citado na página 16.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATY, A.; KHOSLA, A.; BERNSTEIN, M. S.; BERG, A. C.; FEI-FEI, L. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, v. 115, p. 211–252, 2015. Citado na página 28.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. [S.l.]: Pearson, 2016. (Always learning). ISBN 9781292153964. Citado na página 24.

SAMANT, S. S.; CHAUHAN, A.; DN, J.; SINGH, V. Glomerulus detection using segmentation neural networks. **Journal of Digital Imaging**, Springer, v. 36, n. 4, p. 1633–1642, 2023. Citado na página 40.

SANTOS, J. D.; VERAS, R. d. M. S.; VELOSO, R. R.; ALDEMAN, N. L. S.; AIRES, K. R. T.; BIANCHI, A. G. C. et al. Classificação de imagens de biópsias renais com glomeruloesclerose segmentar e focal ou com lesões mínimas utilizando transfer learning em cnn. In: SBC. **Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. [S.l.], 2019. p. 82–93. Citado 2 vezes nas páginas 41 e 43.

SANTOS, J. D.; VERAS, R. de M.; SILVA, R. R.; ALDEMAN, N. L.; ARAÚJO, F. H.; DUARTE, A. A.; TAVARES, J. M. R. A hybrid of deep and textural features to differentiate glomerulosclerosis and minimal change disease from glomerulus biopsy images. **Biomedical Signal Processing and Control**, v. 70, p. 103020, 2021. ISSN 1746-8094. Citado 3 vezes nas páginas 41, 43 e 46.

SANTOS, W. L. dos; FREITAS, L. A. de; DUARTE, A. A.; ANGELO, M. F.; OLIVEIRA, L. R. Computational pathology, new horizons and challenges for anatomical pathology. **Surgical and Experimental Pathology**, BioMed Central, v. 5, n. 1, p. 1–7, 2022. Citado na página 40.

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: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 618–626. Citado na página 36.

SESSO, R. C.; LOPES, A. A.; THOMÉ, F. S.; LUGON, J. R.; MARTINS, C. T. Inquérito brasileiro de diálise crônica 2014. **Brazilian Journal of Nephrology**, SciELO Brasil, v. 38, p. 54–61, 2016. Citado na página 16.

SHEEHAN, S. M.; KORSTANJE, R. Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning. **American Journal of Physiology-Renal Physiology**, American Physiological Society Bethesda, MD, v. 315, n. 6, p. F1644–F1651, 2018. Citado na página 40.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, Springer, v. 6, n. 1, p. 1–48, 2019. Citado 3 vezes nas páginas 48, 49 e 75.

SHUBHAM, S.; JAIN, N.; GUPTA, V.; MOHAN, S.; ARIFFIN, M. M.; AHMADIAN, A. Identify glomeruli in human kidney tissue images using a deep learning approach. **Soft Computing**, Springer, p. 1–12, 2021. Citado na página 40.

SILVA, J.; SOUZA, L.; CHAGAS, P.; CALUMBY, R.; SOUZA, B.; PONTES, I.; DUARTE, A.; PINHEIRO, N.; SANTOS, W.; OLIVEIRA, L. Boundary-aware glomerulus segmentation: Toward one-to-many stain generalization. **Computerized Medical Imaging and Graphics**, Elsevier, v. 100, p. 102104, 2022. Citado na página 40.

SILVA, L.; MACHADO, V.; VERAS, R.; AITA, K.; MONTE, S. do; ALDEMAN, N.; SANTOS, J. Using clinical data and deep features in renal pathologies classification. In: ABRAHAM, A.; PLLANA, S.; CASALINO, G.; MA, K.; BAJAJ, A. (Ed.). **Intelligent Systems Design and Applications**. Cham: Springer Nature Switzerland, 2023. p. 137–146. ISBN 978-3-031-35507-3. Citado na página 17.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: BENGIO, Y.; LECUN, Y. (Ed.). **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 29 e 52.

Sociedade Brasileira de Nefrologia. **Glomerulopatias**. 2023. Acesso em: 26/02/2023. Disponível em: <<https://www.sbn.org.br/orientacoes-e-tratamentos/doencas-comuns/glomerulopatias/>>. Citado na página 20.

SOCIETY, A. C. **What is kidney cancer?** 2017. Disponível em: <<https://www.cancer.org/cancer/kidney-cancer/about/what-is-kidney-cancer>>. Acesso em: 27 de nov. de 2020. Citado na página 16.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCHE, V.; RABINOVICH, A. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 1–9. Citado na página 31.

SZEGEDY, C.; VANHOUCHE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 2818–2826. Citado 2 vezes nas páginas 31 e 52.

TAJBAKHSH, N.; SHIN, J. Y.; GURUDU, S. R.; HURST, R. T.; KENDALL, C. B.; GOTWAY, M. B.; LIANG, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? **IEEE transactions on medical imaging**, v. 35, p. 1299–1312, 2016. Citado 3 vezes nas páginas 28, 50 e 65.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. **International Conference on Machine Learning, 2019**, maio 2019. Citado 2 vezes nas páginas 33 e 52.

TERVAERT, T. W. C.; MOOYAART, A. L.; AMANN, K.; COHEN, A. H.; COOK, H. T.; DRACHENBERG, C. B.; FERRARIO, F.; FOGO, A. B.; HAAS, M.; HEER, E. D. et al. Pathologic classification of diabetic nephropathy. **Journal of the American society of nephrology**, Am Soc Nephrol, v. 21, n. 4, p. 556–563, 2010. Citado na página 41.

TIELEMAN, T.; HINTON, G. et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. **COURSERA: Neural networks for machine learning**, v. 4, p. 26–31, 2012. Citado na página 54.

YI, T. W.; LAING, C.; KRETZLER, M.; NKULIKIYINKA, R.; LEGRAND, M.; JARDINE, M.; ROSSIGNOL, P.; SMYTH, B. Digital health and artificial intelligence in kidney research: a report from the 2020 Kidney Disease Clinical Trialists (KDCT) meeting. **Nephrology Dialysis Transplantation**, v. 37, n. 4, p. 620–627, 11 2021. ISSN 0931-0509. Citado na página 17.

YU, H.; YANG, L. T.; ZHANG, Q.; ARMSTRONG, D.; DEEN, M. J. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. **Neurocomputing**, v. 444, p. 92–110, 2021. ISSN 0925-2312. Citado na página 17.

ZHANG, Q.-s.; ZHU, S.-C. Visual interpretability for deep learning: a survey. **Frontiers of Information Technology & Electronic Engineering**, Springer, v. 19, n. 1, p. 27–39, 2018. Citado na página 36.

ZHENG, Y.; CASSOL, C. A.; JUNG, S.; VEERAPANENI, D.; CHITALIA, V. C.; REN, K. Y.; BELLUR, S. S.; BOOR, P.; BARISONI, L. M.; WAIKAR, S. S.; BETKE, M.; KOLACHALAMA, V. B. Deep-learning–driven quantification of interstitial fibrosis in digitized kidney biopsies. **The American Journal of Pathology**, v. 191, n. 8, p. 1442–1453, 2021. ISSN 0002-9440. Citado na página 40.

ZHENG, Z.; ZHANG, X.; DING, J.; ZHANG, D.; CUI, J.; FU, X.; HAN, J.; ZHU, P. Deep learning-based artificial intelligence system for automatic assessment of glomerular pathological findings in lupus nephritis. **Diagnostics**, v. 11, n. 11, 2021. ISSN 2075-4418. Citado 2 vezes nas páginas 41 e 43.

ZHOU, Z. **Ensemble Methods: Foundations and Algorithms**. [S.l.]: Taylor & Francis, 2012. (CHAPMAN & HALL/CRC MACHINE LEA). ISBN 9781439830031. Citado na página 25.