

Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Classificação Multiclasse do Microambiente Tumoral no Câncer Gástrico Utilizando Ensemble Hierárquico de Dois Níveis com Modelos de Aprendizado Profundo

François Fernandes Ribeiro Barbosa

Teresina-PI, 09 de Fevereiro de 2026

François Fernandes Ribeiro Barbosa

Classificação Multiclasse do Microambiente Tumoral no Câncer Gástrico Utilizando Ensemble Hierárquico de Dois Níveis com Modelos de Aprendizado Profundo

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI, como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Ivan Saraiva Silva

Coorientador: Prof. Dr. Rodrigo de Melo Souza Veras

Teresina-PI

09 de Fevereiro de 2026

Ficha elaborada de acordo com o Código de Catalogação Anglo-Americano AACR2

B238c

Barbosa, François Fernandes Ribeiro.
Classificação multiclasse do microambiente tumoral no câncer gástrico utilizando ensemble hierárquico de dois níveis com modelos de aprendizado profundo. / François Fernandes Ribeiro Barbosa. – Teresina, PI, 2026.
85 f.: il. (alg. color.)

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Programa de Pós-Graduação em Ciência da Computação, 2026.

Orientador: Prof. Dr. Ivan Saraiva Silva.
Coorientador: Prof. Dr. Rodrigo de Melo Souza Veras.

1. Inteligência artificial - Aplicações médicas. 2. Estômago - Tumores. 3. Câncer - Diagnóstico. 4. Aparelho digestivo - Câncer. I. Silva, Ivan Saraiva. II. Veras, Rodrigo de Melo Souza. III. Título.

CDU: 004.8:616.33-006
CDD: 006.31

Ficha elaborada pelo Serviço de Catalogação da Biblioteca Prof. Queirois, IFMA, Campus Caxias
Ianna Torres Lustosa (Bibliotecária) CRB-13/687

François Fernandes Ribeiro Barbosa

Classificação Multiclasse do Microambiente Tumoral no Câncer Gástrico Utilizando Ensemble Hierárquico de Dois Níveis com Modelos de Aprendizado Profundo

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI, como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 09 de Fevereiro de 2026:

gov.br Documento assinado digitalmente
IVAN SARAIVA SILVA
Data: 17/03/2026 11:10:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ivan Saraiva Silva
Orientador

gov.br Documento assinado digitalmente
RODRIGO DE MELO SOUZA VERAS
Data: 17/03/2026 16:14:18-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo de Melo Souza Veras
Co-Orientador

gov.br Documento assinado digitalmente
KELSON ROMULO TEIXEIRA AIRES
Data: 18/03/2026 17:10:04-0300
Verifique em https://validar.iti.gov.br

Prof. Dr Kelson Romulo Teixeira Aires
Convidado 1

gov.br Documento assinado digitalmente
VINICIUS PONTE MACHADO
Data: 19/03/2026 11:45:27-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Vinicius Ponte Machado
Convidado 2

gov.br Documento assinado digitalmente
MARIANA RECAMONDE MENDOZA GUERREIRO
Data: 19/03/2026 13:06:00-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Mariana Recamonde Mendoza
Convidado 3

Teresina-PI
09 de Fevereiro de 2026

À minha esposa, por sempre estar comigo em todos os momentos, com amor, apoio e companheirismo.

Agradecimentos

Agradeço a Deus, por sua infinita misericórdia, por acolher minhas infidelidades e, ainda assim, conceder-me bênçãos e conquistas que ultrapassam o que considero merecer.

Agradeço à minha esposa, Iris Barbosa, e aos meus filhos, Isaac, Clarice, Maria Liz e Lucas Barbosa, por serem minha base e minha força diária, sustentando-me com amor, paciência e incentivo para que eu pudesse buscar e concluir esta conquista.

Agradeço aos meus pais, pelo cuidado, pela dedicação e pela entrega de tantos anos, fundamentais para que eu chegasse até aqui.

Agradeço ao meu orientador, Professor Ivan Saraiva, pela seriedade, sinceridade e humanidade ao conduzir esta orientação, contribuindo de forma decisiva para meu crescimento acadêmico e pessoal.

Agradeço aos meus amigos de turma, em especial Isaac Ramos, Fernando Jorge e Jeovane Cordeiro, pela parceria, companheirismo e apoio ao longo desta trajetória, especialmente nos momentos mais desafiadores.

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA) Campus Caxias, por proporcionar as condições necessárias para o desenvolvimento deste trabalho e para a realização desta etapa formativa.

Por fim, agradeço a todos os professores do programa, pelas contribuições, ensinamentos e dedicação, que foram essenciais para minha formação e para a conclusão deste percurso.

Resumo

O câncer gástrico (CG) permanece entre as principais causas de morte por câncer no Brasil e no mundo, com altas taxas de incidência e mortalidade, especialmente em países em desenvolvimento. A oncologia moderna evoluiu de modo a estratificar os pacientes em grupos distintos, com base nas características genotípicas e fenotípicas dos tumores, bem como em seus respectivos desfechos clínicos. Nesse contexto, o estudo do microambiente tumoral (TME) é essencial para a definição de condutas terapêuticas individualizadas e eficazes. Assim, a classificação automatizada dos componentes do TME por meio de modelos de aprendizado profundo (DL *Deep Learning*) surge como uma alternativa promissora. Apesar dos avanços em patologia digital, a classificação automática dos diferentes componentes teciduais do TME em lâminas coradas com *Hematoxilina e Eosina* (H&E) ainda é pouco explorada em CG. Diante dessa lacuna, este trabalho propõe um método de classificação multiclasse, baseado em um *ensemble* hierárquico de dois níveis, composto por modelos de DL. Foram avaliadas doze arquiteturas pré-treinadas em ImageNet (EfficientNet, ViT, MobileNet, Xception, VGG, ResNet, DenseNet, AlexNet, entre outras), explorando transferência de aprendizado, extração e concatenação de *embeddings*, arranjos hierárquicos de classificadores binários, variação do número de neurônios nas camadas totalmente conectadas e técnicas de aumento de dados em tempo de execução. Os experimentos foram conduzidos sobre o conjunto de dados HMU-GC-HE-30K, composto por 31.096 *patches* balanceados em oito classes de tecidos, utilizando validação cruzada estratificada ($k = 8$), e 20% das imagens reservadas para teste externo. A partir dos melhores modelos, foi construído um *ensemble* hierárquico de dois níveis: comitês locais homogêneos e, em seguida, um comitê global entre arquiteturas distintas, usando as predições do comitê local. Os melhores resultados obtidos nos conjuntos de teste indicaram que comitê mais performático obteve acurácia de 81,2%, AUC de 98,1% e *F1-score* de 81,2%, superando em mais de dez pontos percentuais o *baseline* da literatura para essa base, e demonstrando a superioridade do *ensemble* hierárquico frente a classificadores isolados. Esse comitê usou como estratégia de apuração a média ponderada pelo *F1-score* de cada classe de classificador membros. Esses achados mostram que a combinação estruturada de arquiteturas diversas, aplicada a um comité hierárquico, reduz a variância entre execuções e aumenta a capacidade discriminativa em cenários multiclasse complexos, contribuindo para o avanço da patologia digital assistida por Inteligência Artificial (IA).

Palavras-chaves: Microambiente tumoral, Câncer gástrico, Aprendizado profundo, Redes neurais convolucionais, Ensemble hierárquico.

Abstract

Gastric cancer (GC) remains one of the leading causes of cancer-related death in Brazil and worldwide, with high incidence and mortality rates, especially in developing countries. Modern oncology has evolved toward stratifying patients into distinct groups based on the genotypic and phenotypic characteristics of tumors, as well as on their respective clinical outcomes. In this context, the study of the tumor microenvironment (TME) is essential for defining individualized and effective therapeutic strategies. Thus, the automated classification of TME components using deep learning (DL) models emerges as a promising alternative. Despite advances in digital pathology, the automatic classification of different TME tissue components in hematoxylin and eosin (H&E)stained slides remains underexplored in GC. To address this gap, this work proposes a multiclass classification method based on a two-level hierarchical ensemble composed of DL models. Twelve ImageNet-pretrained architectures (EfficientNet, ViT, MobileNet, Xception, VGG, ResNet, DenseNet, AlexNet, among others) were evaluated, exploring transfer learning, embedding extraction and concatenation, hierarchical arrangements of binary classifiers, variation in the number of neurons in fully connected layers, and on-the-fly data augmentation techniques. The experiments were conducted on the HMU-GC-HE-30K dataset, which comprises 31,096 patches balanced across eight tissue classes, using stratified cross-validation ($k = 8$), with 20% of the images held out for external testing. Based on the best-performing models, we built a two-level hierarchical ensemble: homogeneous local committees and, subsequently, a global committee across distinct architectures using the predictions of the local committees. The best results obtained on the test sets indicated that the highest-performing committee achieved an accuracy of 81.2%, an AUC of 98.1%, and an F1-score of 81.2%, outperforming the literature baseline for this dataset by more than ten percentage points and demonstrating the superiority of the hierarchical ensemble over individual classifiers. This ensemble used as its aggregation strategy a class-wise F1-scoreweighted average of the member classifiers predictions. These findings show that the structured combination of diverse architectures within a hierarchical ensemble reduces variance across runs and increases discriminative capacity in complex multiclass scenarios, contributing to the advancement of artificial intelligence (AI)assisted digital pathology.

Keywords: Tumor microenvironment, Gastric cancer, Deep learning, Convolutional neural networks, Hierarchical ensemble.

Lista de ilustrações

Figura 1 – Ilustração da incidência/mortalidade do CG no Mundo. Em (a) incidência de casos e em (b) quantidade de mortos por CG. Fonte: (FERLAY et al., 2024) . . . 15
Figura 2 – Ilustração da incidência do CG no mundo por concentração. Fonte: (FERLAY et al., 2024) . . . 16
Figura 3 – Exemplo de (a) lâmina histológica do TME de CG e (b) mosaico de *patches* dos diversos componentes do TME. Fonte: (LOU et al., 2025) . 16
Figura 4 – Anatomia apresentando (a) as partes do estômago e (b) as camadas da parede do estômago. Fonte: adaptado de <https://www.cancer.org> . 20
Figura 5 – Ilustração de tecido gástrico corado com H&E. Fonte: (LOU et al., 2024) 21
Figura 6 – Ilustração da organização hierárquica da IA. Fonte: Autor . . . 22
Figura 7 – Arquitetura genérica de CNNs. Fonte: Inspirado em (LECUN et al., 1998) . . . 24
Figura 8 – Arquitetura da VGGNet com 16 camadas . . . 25
Figura 9 – Arquitetura da rede ResNet101, composta por uma camada convolucional inicial, seguida por blocos residuais e finalizada por *global average pooling* e uma camada totalmente conectada. . . . 26
Figura 10 – Esquema de um bloco denso da DenseNet, em que cada camada recebe como entrada a concatenação das saídas de todas as camadas anteriores. 26
Figura 11 – Esquema comparativo entre (a) convolução 2D padrão, (b) convolução *depthwise* e (c) convolução 1×1 (*pointwise*), destacando a relação entre volume de entrada, filtros e volume de saída em cada caso. . . . 27
Figura 12 – Arquitetura da rede EfficientNet_B0 mostrando apenas as camadas convolucionais e o *global average pooling*. . . . 28
Figura 13 – Arquitetura geral do Vision Transformer (ViT), da divisão da imagem em *patches* ao codificador *Transformer* e à MLP de classificação. Fonte: Inspirado em Dosovitskiy et al. (2021). . . . 29
Figura 14 – Ilustração de uma arquitetura comum de *ensemble*. Fonte: Inspirado em (ZHOU, 2012) . . . 30
Figura 15 – Exemplo de curva ROC, destacando diferentes limiares de decisão e a área sob a curva (AUC). . . . 33
Figura 16 – Exemplo da *string* de busca utilizada na base da Pubmed. . . . 34
Figura 17 – Fluxo dos experimentos. . . . 40
Figura 18 – *Patches* extraídos rotulados com as oito classes de tecido TME. Fonte: (LOU et al., 2025) . . . 42

Figura 19 – Distribuição de validação cruzada com k=8 e 20% do total dos dados separados para teste. 43
Figura 20 – Distribuição das imagens no conjunto de treino utilizando agrupamento intra-classe. À esquerda, o *dataset* é dividido em *k folds*. À direita, exemplo da classe MUS: cada ponto representa um *patch* de imagem, colorido de acordo com o cluster ao qual pertence. 44
Figura 21 – Diferentes estratégias de extração de características utilizando operações de GAP. (a) Arquitetura Original;(b) Estratégia 4 GAPs; (c) inclusão de um GAP adicional após a primeira camada convolucional; (d) versão simplificada com apenas os dois últimos GAPs 46
Figura 22 – Agrupamento hierárquico das classes, apresentando baixo nível de conflito (lado A) e alto nível de conflito (lado B) em relação ao tecido tumoral. 47
Figura 23 – Exemplo de fluxo de decisão hierárquica com classificadores binários, resultando na classe final. 47
Figura 24 – Arquiteturas com modificações na camada totalmente conectada diminuindo a quantidade de neurônios (256 e 512). 48
Figura 25 – Imagens dos aumentos utilizados. a) Imagem original seguidas de transformações de rotação, espelhamentos horizontal e vertical, inversão diagonal e anti-diagonal. b) Imagem original seguidas de translação, rotação, *zoom out, zoom in* e cisalhamento; c) Imagem original seguida variações de cor; d) imagem original seguida de composições diferentes de blocos embaralhados 49
Figura 26 – Representação do *ensemble* local. Cada um dos oito classificadores base, derivados das diferentes execuções da validação cruzada, gera uma predição individual que são então combinadas, resultando na predição local do comitê. 51
Figura 27 – Estrutura do *ensemble* global. As predições dos *ensembles* locais são combinadas para gerar a predição final. 52
Figura 28 – Exemplo da Estratégia I de apuração: as probabilidades previstas pelos classificadores para cada classe são combinadas pela média aritmética, produzindo o vetor de probabilidades finais do *ensemble*. 53
Figura 29 – Exemplo da Estratégia II de apuração: as predições discretas de cada classificador são somadas por classe (votação) e, em seguida, normalizadas para gerar o vetor de probabilidades finais do *ensemble*. 54
Figura 30 – Exemplo da Estratégia III de apuração: para cada classificador, apenas a maior probabilidade é mantida (demais tornam-se zero) e, em seguida, essas probabilidades são somadas e normalizadas para compor o vetor de probabilidades finais do *ensemble*. 54

Figura 31 – Exemplo da Estratégia IV de apuração: as probabilidades previstas por cada classificador são combinadas por média ponderada, utilizando pesos normalizados baseados no *F1-score* de cada modelo. 55
Figura 32 – Exemplo da Estratégia V de apuração: os votos discretos de cada classificador são ponderados pelos pesos normalizados de *F1-score* global e somados por classe, gerando as probabilidades finais do *ensemble*. . . 55
Figura 33 – Exemplo da Estratégia VI de apuração: para cada classificador, apenas a maior probabilidade é mantida e ponderada pelos pesos normalizados de *F1-score* global; em seguida, essas contribuições são somadas por classe para formar a predição final do *ensemble*. 56
Figura 34 – Exemplo da Estratégia VII de apuração: as probabilidades previstas por cada classificador são ponderadas pelos pesos específicos de *F1-score* por classe e, em seguida, normalizadas para gerar o vetor de probabilidades finais do *ensemble*. 56
Figura 35 – Exemplo da Estratégia VIII de apuração: os votos discretos de cada classificador são ponderados pelos pesos específicos de *F1-score* por classe e, após normalização, formam o vetor de probabilidades finais do *ensemble*. 57
Figura 36 – Exemplo da Estratégia IX de apuração: para cada classificador, apenas a maior probabilidade é mantida, ponderada pelos pesos específicos de *F1-score* por classe e, em seguida, normalizada para compor o vetor de probabilidades finais do *ensemble*. 57
Figura 37 – Projeção bidimensional dos GAPs antes da concatenação. a) Vetores após a primeira convolução; b) apos a segunda convolução; c) após a quarta convolução; d) após a sexta convolução e e) após todas as convolução. 62
Figura 38 – Gráficos de radar comparando as métricas F1-score (a), Precisão (b) e Recall (c) entre o modelo proposto e os resultados dos autores da base. 71
Figura 39 – Comparação entre as matrizes de confusão dos modelos replicado e proposto, destacando melhorias na separabilidade entre classes comumente confundidas. 72
Figura 40 – *Loss* e acurácia das redes do ensemble c20 em cada época nos conjuntos de validação e treino 73
Figura 41 – Mapas ativação de exemplos das oito classes de tecido para os cinco modelos avaliados. A primeira coluna mostra a imagem original e a barra superior indica acertos (verde) e erros (vermelho). 75

Lista de tabelas

Tabela 1 – Síntese comparativa dos trabalhos relacionados. 37
Tabela 2 – Lista das arquiteturas avaliadas. 45
Tabela 3 – Resumo dos hiperparâmetros testados. 50
Tabela 4 – Descrição dos métodos de combinação de classificadores utilizados. . . 52
Tabela 5 – Resultados do estudo de Lou et al. (2025) e da replicação deste estudo. 59
Tabela 6 – Resultados comparativos dos modelos sem alterações na arquitetura. . 60
Tabela 7 – Resultados de modificações arquiteturais usando GAPs concatenados. . 61
Tabela 8 – Resultados comparativos do uso de classificadores hierárquicos 63
Tabela 9 – Resultados do candidatos a membros do comitê. 63
Tabela 10 – Resultados comparativos dos modelos com variação na quantidade de neurônios na camada totalmente conectada. 64
Tabela 11 – Resultados dos modelos com e sem aumento de dados. 66
Tabela 12 – Resultados dos *ensembles* locais usando apuração por média e por votação, a partir das predições com aumento de dados 67
Tabela 13 – Combinações de classificadores e suas respectivas abreviações. 68
Tabela 14 – Resultados comparativos das combinações de classificadores. 69
Tabela 15 – Resultados comparativos entre *Ensemble* tradicional e o *Ensemble* proposto. 70
Tabela 16 – Resultados comparativos entre os autores da base e o modelo proposto. 70
Tabela 17 – Resultados comparativos por classe entre os autores da base e o modelo proposto (ganho relativo em pontos percentuais). 70
Tabela 18 – Resultados com variação na quantidade de neurônios dos classificadores binários, sem aumento de dados. 84
Tabela 19 – Resultados com dos melhores classificadores binários com aumento de dados. 84
Tabela 20 – Resultados cos classificadores de cada um dos níveis hierárquicos. . . . 84

Lista de abreviaturas e siglas

alx AlexNet

AUC Área sob a Curva

CG Câncer Gástrico

CNN Rede Neural Convolucional (*Convolutional Neural Network*)

CRC Câncer Colorretal (*Colorectal Cancer*)

DACHS *Darmkrebs: Chancen der Verhütung durch Screening*

DL Aprendizado Profunda (*Deep Learning*)

ROC *Receiver Operating Characteristic*

efn EfficientNet-B0

FN Falsos Negativos

FP Falsos Positivos

FPR Taxa de Falsos Positivos (*False Positive Rate*)

GAP *Global Average Pooling*

GCO Observatório Global de Câncer (*Global Cancer Observatory*)

Grad-CAM *Gradient-weighted Class Activation Map*

H&E Hematoxilina e Eosina

IA Inteligência Artificial

INCA Instituto Nacional de Câncer

IRQ Imunoistoquímica (*Immunohistochemistry*)

LR Taxa de Aprendizado (*Learning Rate*)

mbn MobileNetV3

ML Aprendizado de Máquina (*Machine Learning*)

NLST *National Lung Screening Trial*

NuCLS *Nuclear Classification and Labeling System*

OST *On-Step Tuning*

RNA Rede Neural Artificial

rsn101 ResNet101

rsn50 ResNet50

TCGA *The Cancer Genome Atlas*

t-SNE *t-distributed Stochastic Neighbor Embedding*

TL *Transfer Learning*

TME Microambiente Tumoral (*Tumor Microenvironment*)

TPR Taxa de Verdadeiros Positivos (*True Positive Rate*)

VN Verdadeiros Negativos

VP Verdadeiros Positivos

vit *Vision Transformer*

vgg VGG16

WSI *Whole Slide Imaging*

xcp Xception

Sumário

1 INTRODUÇÃO . . . 15
1.1 Objetivo Geral . . . 17
1.2 Objetivos Específicos . . . 18
1.3 Organização do Trabalho . . . 18

2 FUNDAMENTAÇÃO TEÓRICA . . . 19
2.1 Câncer Gástrico . . . 19
2.2 Microambiente Tumoral (TME) . . . 20
2.3 Histopatologia Digital . . . 20
2.4 Inteligência Artificial . . . 21
2.4.1 Aprendizado de Máquina . . . 23
2.4.2 Redes Neurais Artificiais . . . 23
2.4.3 Redes Neurais Convolucionais . . . 24
2.4.3.1 VGG . . . 24
2.4.3.2 Resnet . . . 25
2.4.3.3 DenseNet . . . 26
2.4.3.4 MobileNet . . . 27
2.4.3.5 *Efficientnet* . . . 28
2.4.3.6 *Vision Transformer* . . . 28
2.5 Aumento de Dados . . . 29
2.6 Comitê de Classificadores . . . 30
2.7 Métricas de avaliação . . . 31
2.8 Considerações Finais . . . 33

3 TRABALHOS RELACIONADOS . . . 34
3.1 Estratégia de busca e seleção dos estudos . . . 34
3.2 Estudos selecionados . . . 34
3.3 Posicionamento deste trabalho na literatura . . . 37
3.4 Considerações Finais . . . 38

4 MATERIAIS E MÉTODOS . . . 40
4.1 Preparação dos dados . . . 41
4.1.1 Conjuntos de dados . . . 41
4.1.2 Validação Cruzada . . . 42
4.2 Modelagem dos Modelos . . . 44
4.2.1 Fase 1: Experimentos com Arquiteturas Originais . . . 44

4.2.2 Fase 2: Experimentos com Foco nos Vetores de Características 45
4.2.3 Fase 3: Experimentos com Hierarquização de treinamentos 46
4.2.4 Fase 4: Experimentos Focados em Quantidade de Neurônios 48
4.2.5 Aumento de dados 48
4.3 Estratégias de treinamento 50
4.4 Formação de Comitês 50
4.4.1 Ensemble Local 51
4.4.2 Ensemble Global 51
4.4.3 Estratégia de Apuração 52
4.5 Reprodutibilidade 57
4.6 Considerações Finais 57

5 RESULTADOS E DISCUSSÃO 59
5.1 Resultados do Estudo Original 59
5.2 Resultados com Arquitetura original 60
5.3 Resultados com Foco nos Vetores de Características 61
5.4 Resultados com Hierarquização de treinamentos 62
5.5 Resultados Focados em Quantidade de Neurônios 63
5.6 Resultados com *Ensemble* de dois Níveis 65
5.6.1 Nível 1: *Ensemble* Local 66
5.6.2 Nivel 2: *Ensemble* Global 67
5.7 Explicabilidade 74
5.8 Considerações Finais 74

6 CONCLUSÕES E TRABALHOS FUTUROS 76

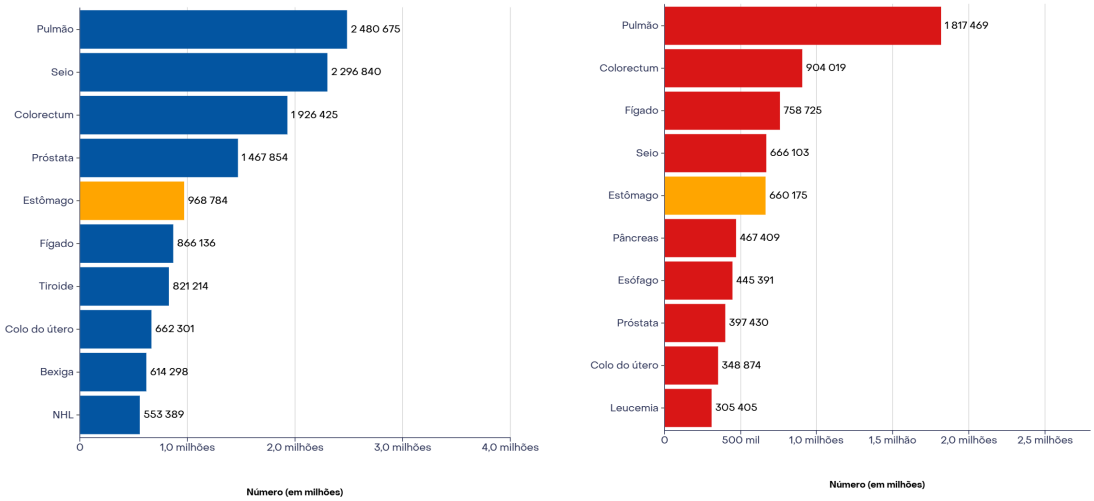
REFERÊNCIAS 78

ANEXOS 83

ANEXO A – RESULTADOS DOS CLASSIFICADORES BINÁRIOS 84

1 Introdução

O Cancêr Gástrico (CG) ocupa a quinta posição entre os adenocarcinomas com maior incidência no mundo (Figura 1-a), com mais de 960 mil casos registrados (representando 5,96% de todos os casos de câncer), somente em 2022, além de ser o quinto colocado entre as principais causas de morte por câncer (Figura 1-b), ultrapassando as 660 mil mortes por ano (6,82% de todas as mortes relacionadas a câncer), de acordo com os dados do Relatório Global de Estatísticas do Câncer de 2022 do Observatório Global de Câncer (*Global Cancer Observatory* GCO) (FERLAY et al., 2024).



(a) (b)

Figura 1 – Ilustração da incidência/mortalidade do CG no Mundo. Em (a) incidência de casos e em (b) quantidade de mortos por CG. Fonte: (FERLAY et al., 2024)

O continente asiático concentra a maior incidência de novos casos de CG, seguido da América do Sul e do Caribe, como podemos verificar na Figura 2. No Brasil, esse tipo de câncer ocupa o sexto lugar em incidência, sendo a região Norte a mais impactada, com o estado do Amazonas apresentando a maior concentração de casos por 100 mil habitantes, segundo o INCA (Instituto Nacional de Câncer (INCA), 2024).

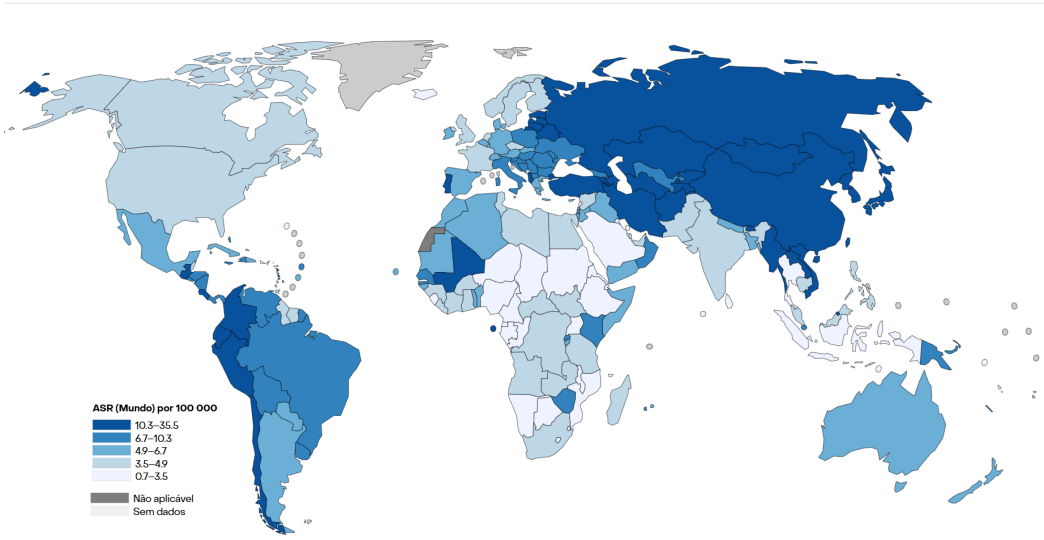


Figura 2 – Ilustração da incidência do CG no mundo por concentração. Fonte: (FERLAY et al., 2024)

O entendimento atual sobre câncer vai além da visão de uma mera doença genética, sendo compreendido como um ecossistema complexo incluindo uma variedade de células não cancerosas que interagem com o tumor, formando o chamado microambiente tumoral (TME - *tumor microenvironment*) (Figura 3-a) (de Visser; JOYCE, 2023). Esse microambiente desempenha um papel importante no desenvolvimento da doença, assim como na resposta terapêutica. Nesse sentido, há uma necessidade clínica de classificar, entre os diversos componentes do TME (Figura 3-b), as estruturas presentes em imagens histopatológicas de pacientes com CG (LOU et al., 2025).

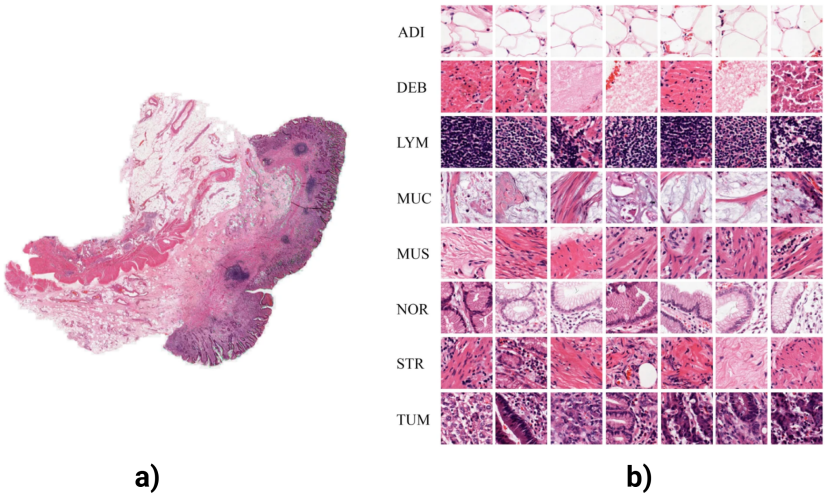


Figura 3 – Exemplo de (a) lâmina histológica do TME de CG e (b) mosaico de *patches* dos diversos componentes do TME. Fonte: (LOU et al., 2025)

O diagnóstico tradicional do CG é feito por meio de análise visual de cortes histopatológicos do tecido gástrico corados com *Hematoxilina e Eosina* (H&E). No entanto,

esse método é trabalhoso, demorado e subjetivo, pois depende da experiência do patologista. Esses fatores comprometem a acurácia diagnóstica especialmente diante da escassez de patologistas experientes e alta carga de trabalho na patologia diagnóstica (WANG et al., 2024).

A patologia é uma área da medicina responsável pelo estudo dos órgãos e tecidos do corpo humano, sendo fundamental para o diagnóstico de diversas doenças. Contudo, com os avanços recentes, métodos tradicionais de diagnóstico têm dado espaço para uso da inteligência artificial (IA) em virtude das digitalizações de lâminas inteiras (WSI - *Whole Slide Imaging*) e na patologia digital (SERAG et al., 2019).

A possibilidade de gerar imagens digitais de alta qualidade, capazes de capturar padrões morfológicos complexos, traz inúmeras possibilidades para o uso de IA na detecção de doenças, especialmente no diagnóstico de câncer e na análise de biomarcadores teciduais. Essa área, embora ainda em estágio inicial, apresenta grande potencial para transformar o estudo de patologias, aumentando a precisão nos diagnósticos e melhorando os desfechos clínicos dos pacientes (SERAG et al., 2019).

Nas últimas décadas, observa-se um expressivo crescimento na utilização de algoritmos de Aprendizado de Máquina (*ML - Machine Learning*) em pesquisas oncológicas, especialmente aqueles baseados em Redes Neurais Convolucionais (*CNN - Convolutional Neural Network*), onde se tem excelentes resultados na análise de imagens digitalizadas e no suporte ao diagnóstico clínico (MANDAL et al., 2025). Nesse contexto, a capacidade dessas redes em extrair, das imagens, características morfológicas relevantes, superam limitações subjetivas da análise manual contribuindo para maior acurácia diagnóstica.

A identificação dos diferentes componentes do TME, em imagens histológicas do CG é de fundamental importância, tendo em vista que essas estruturas desempenham papéis cruciais tanto na progressão tumoral como na resposta ao tratamento. Frente ao crescente interesse da patologia moderna em explorar esses elementos, este trabalho propõe o desenvolvimento de um comitê de modelos baseado em aprendizado profundo (DL - *Deep Learning*) para a classificação dos diversos tipos de tecido que compõem o TME no contexto do CG.

1.1 Objetivo Geral

O objetivo deste trabalho é desenvolver um comitê (*ensembles*) de redes neurais, em dois níveis, para a classificação multiclasse de imagens histológicas de tecidos do TME de CG, coradas com H&E, combinando predições para identificar as diferentes estruturas morfológicas que compõem o TME.

1.2 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, propõe-se os seguintes objetivos específicos:

- Investigar e avaliar as principais arquiteturas de DL apresentadas na literatura para classificação de imagens do TME;
- Avaliar diferentes estratégias de treinamento e configuração dos modelos, testá-las individualmente e analisar seus desempenhos com base em métricas específicas de classificação multiclasse;
- Aplicar técnicas de aumento de dados (*data augmentation*) sobre as imagens de treinamento, com o objetivo de aumentar a variabilidade das amostras, buscando melhorar a generalização dos modelos;
- Definir critérios objetivos para a seleção das melhores arquiteturas individuais, visando à construção de classificadores baseados em comitês (*ensembles*) com potencial superior de desempenho;
- Propor e construir um modelo com dois níveis de comitês, sendo um local para cada arquitetura selecionada e um global entre essas arquiteturas.
- Aplicar técnicas de explicabilidade, como o Grad-CAM, para analisar as regiões de interesse utilizadas pelos modelos na tomada de decisão, contribuindo para a interpretação dos resultados e para a confiabilidade do sistema proposto.

1.3 Organização do Trabalho

Os demais capítulos deste trabalho foram organizados da seguinte forma:

- O Capítulo 2 trata da fundamentação teórica necessária para a construção desta pesquisa.
- O Capítulo 3 apresenta uma breve revisão dos trabalhos relacionados à TME e classificação multiclasse de imagens médicas.
- O Capítulo 4 os etapas dos experimentos realizados para se chegar ao método proposto.
- O Capítulo 5 apresenta os resultados obtidos com os experimentos realizados.
- O Capítulo 6 apresenta as considerações finais.

2 Fundamentação Teórica

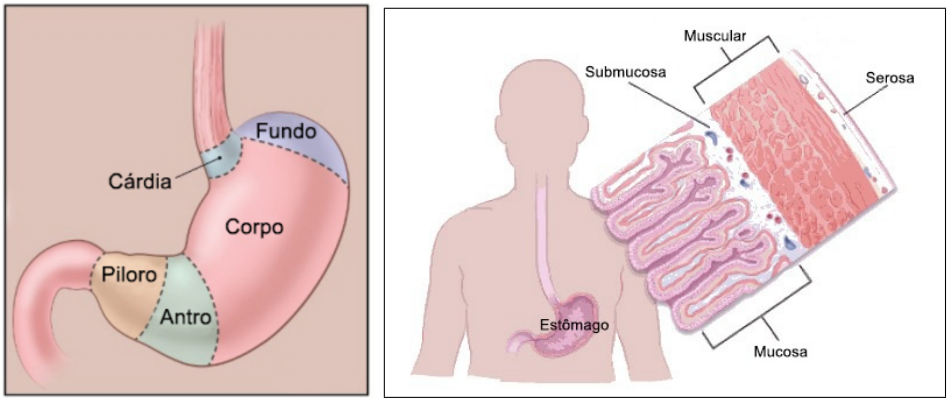
Neste capítulo, são apresentados os conceitos fundamentais para a compreensão do tema desenvolvido neste trabalho. Inicialmente, abordam-se aspectos médicos essenciais, como a estrutura do estômago e o câncer gástrico (Seção 2.1), o microambiente tumoral (Seção 2.2) e a histopatologia digital (Seção 2.3). Na sequência, discutimos os pilares computacionais que sustentam a proposta inteligência artificial, aprendizado de máquina e aprendizado profundo (Seção 2.4) além de técnicas de aumento de dados (Seção 2.5) e comitês de classificadores (Seção 2.6). Por fim, descrevemos as métricas de avaliação (Seção 2.7) e encerramos com as considerações finais (Seção 2.8).

2.1 Câncer Gástrico

O estômago humano é um dos órgãos do sistema digestivo que está localizado entre o esôfago e o intestino delgado. Ele é um órgão grande e oco, que atua como um misturador. Entre suas funções primárias incluem o armazenamento temporário e a digestão química e mecânica dos alimentos. Sua anatomia é categorizada em cinco partes: cárdia, fundo, corpo, antro e piloro, conforme podemos observar na Figura 4-a.(SHAIKH; WEHRLE; KHORASANI-ZADEH, 2023).

A parede do estômago (Figura 4-b) tem em sua composição quatro camadas: mucosa, submucosa, muscular externa e serosa. A mucosa é a camada mais interna, revestida pelo tecido epitelial, compreendido principalmente por glândulas gástricas responsáveis pela secreção do suco gástrico. A submucosa, localizada abaixo da mucosa, é composta por tecido conjuntivo que dá suporte estrutural e contém vasos sanguíneos e linfáticos. A camada muscular externa reveste a submucosa e é responsável pela movimentação da parede do estômago, realizando assim a misturação e propulsão do alimento até o intestino. A serosa é a camada mais externa, constituída por tecido conjuntivo revestido por mesotélio, que reduz o atrito entre o estômago e os órgãos adjacentes (SHAIKH; WEHRLE; KHORASANI-ZADEH, 2023).

Amostras de mucosa gástrica (Figura 3-a) obtidas por biópsia durante exames endoscópicos permitem identificar as principais alterações patológicas, como inflamações, infecções e neoplasias. A mucosa gástrica pode ser retirada de diferentes regiões, como a mucosa oxíntica (fundo/corpo) e a mucosa antral (antro e piloro), cada uma com características histológicas distintas. A escolha do local para biópsia depende da suspeita clínica e das alterações observadas durante a endoscopia (RUGGE et al., 2021).



(a) (b)

Figura 4 – Anatomia apresentando (a) as partes do estômago e (b) as camadas da parede do estômago. Fonte: adaptado de <https://www.cancer.org>

2.2 Microambiente Tumoral (TME)

A oncologia moderna evoluiu de modo a estratificar os pacientes em grupos distintos, com base nas características genotípicas e fenotípicas dos tumores, bem como em seus respectivos desfechos clínicos (KATHER et al., 2019). Nesse contexto, torna-se fundamental ampliar o olhar para além do tecido tumoral em si e considerar todo o contexto em seu entorno, reconhecendo a complexidade e a influência do TME na análise da biologia do câncer.

Este TME é composto por uma variedade de componentes não cancerígenos presentes no interior e ao redor do tumor, tais como células imunes, fibroblastos associados ao câncer, células endoteliais, pericitos e outros tipos de células que variam de acordo com o tecido analisado (de Visser; JOYCE, 2023). Esse ecossistema complexo tem um papel fundamental tanto na progressão tumoral quanto na resposta ao tratamento do câncer. A compreensão do grau de infiltração de células imunes no TME tem se mostrado essencial para a personalização de abordagens terapêuticas, contribuindo significativamente para a redução da mortalidade por câncer (WANG et al., 2024).

2.3 Histopatologia Digital

A patologia é o campo da medicina responsável pelo diagnóstico de doenças por meio da análise de biópsias ou através de procedimentos cirúrgicos. Dentro desta área, existe a histopatologia, que se concentra em analisar amostras microscópicas de tecidos, com o objetivo de diagnosticar doenças específicas, sendo a base para o diagnóstico do CG (MANDAL et al., 2025).

Para uma melhor visualização da arquitetura tecidual, utiliza-se uma técnica de

coloração que possibilita a identificação de anormalidades e doenças. A *Hematoxilina e Eosina* (H&E) é o método de coloração mais utilizado na histopatologia, sendo a primeira responsável por corar os núcleos celulares em tons de azul-escuro, e a segunda, por corar as demais estruturas em tons de rosa (Figura 5)(HOLZINGER et al., 2020)

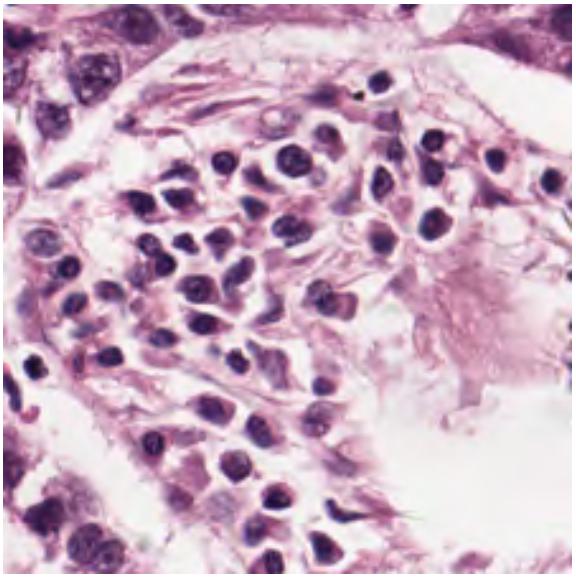


Figura 5 – Ilustração de tecido gástrico corado com H&E. Fonte: (LOU et al., 2024)

A patologia digital é uma das áreas que mais se desenvolveram nos últimos anos, impulsionada pelo desenvolvimento tecnológico e pela demanda por diagnósticos mais precisos. Situada entre a matemática aplicada e a medicina, ela dedica-se ao estudo e ao uso de algoritmos capazes de analisar automaticamente imagens histológicas, com o objetivo de oferecer suporte às decisões clínicas dos patologistas nos diagnósticos de diversas doenças (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022).

O principal objeto de investigação da patologia digital são as imagens de lâminas inteiras, que são obtidas por meio de microscópios digitais de alta performance. Para a análise dessas imagens, destacam-se os métodos baseados em técnicas de aprendizado de máquina, especialmente os modelos de aprendizado profundo, cuja abordagem será detalhada no próximo tópico (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022).

2.4 Inteligência Artificial

A Inteligência Artificial (IA) é uma das áreas mais recentes da ciência, tendo surgido na década de 1950. Segundo Russel e Norvig (RUSSELL; NORVIG, 2016), a IA pode ter definições agrupadas em quatro grupos: pensamento humano, pensamento racional, comportamento humano e comportamento racional. Seu foco principal é o estudo de agentes capazes de perceber e analisar o ambiente ao seu redor para tomada de decisões que maximizem as chances de sucesso. Já Luger (LUGER, 2013) por entender que IA

é um ramo da computação, a define como uma área que se ocupa da automação do comportamento inteligente.

Em 1950, o renomado cientista britânico Alan Turing definiu o que ficou conhecido como o teste de Turing. Para que uma máquina seja considerada inteligente, ela teria que ter a capacidade de responder a perguntas de uma pessoa sem que esta consiga diferenciar se a resposta foi dada por uma máquina ou por outro ser humano (TURING, 1950). Para passar neste teste, a máquina precisaria ter habilidades nas principais áreas da IA, incluindo Processamento de Linguagem Natural, Representação de Conhecimento, Raciocínio Automatizado, Aprendizado de Máquina e Robótica.

Dentre as subáreas mencionadas acima, destacam-se abordagens baseadas em Aprendizado de Máquina (*machine learning* ML), tais como o Aprendizado Profundo (*Deep Learning*-DL), Redes Neurais Convolucionais (*Convolutional Neural Network* -CNN) e *Transformers*, atualmente as mais conhecidas e amplamente utilizadas no auxílio a diagnóstico médico. Na Figura 6 ilustra como essas abordagens se organizam dentro do campo da IA.

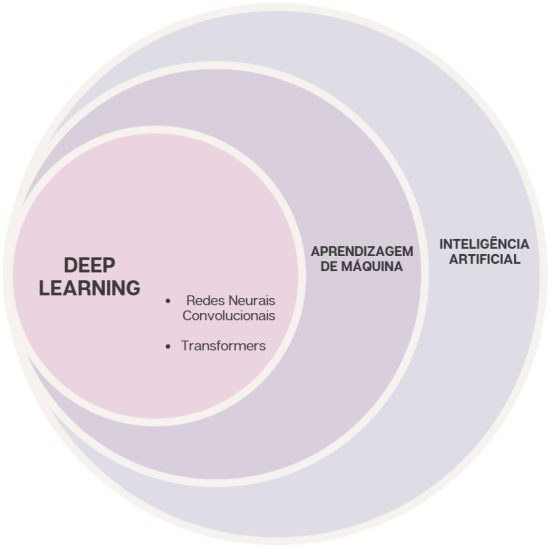


Figura 6 – Ilustração da organização hierárquica da IA. Fonte: Autor

A visão computacional é uma área da ciência da computação, que tem a capacidade de criar descrições claras e dar significado a objetos a partir de imagens obtidas por câmeras, sensores ou qualquer dispositivo capaz de capturar imagens. É importante salientar que o processamento de imagens é responsável por realizar transformações nas imagens, enquanto a visão computacional tem o objetivo de extrair informações de alto nível, correspondentes ao conteúdo semântico das imagens (BALLARD; BROWN, 1982). Em muitos problemas práticos, a visão computacional se apoia em técnicas de inteligência artificial, embora não dependa exclusivamente delas.

2.4.1 Aprendizado de Máquina

O aprendizado de máquina, é o campo da inteligência artificial que, segundo Miller e Guido (GUIDO; MüLLER, 2017), tem como princípio o desenvolvimento de algoritmos que, a partir de dados de entrada, conseguem aprender e tomar decisões. Trata-se de um subcampo da IA que busca a integração entre matemática, estatística e algoritmos computacionais. Esses algoritmos possibilitam a construção de modelos matemáticos cujos parâmetros são otimizados com o objetivo de prever resultados específicos, o que é conhecido como treinamento do modelo. Para que esse processo tenha sucesso, é necessário um amplo conjunto de dados de entrada, de modo que, após os parâmetros estarem bem ajustados, o modelo consiga realizar previsões corretas mesmo quando aplicado a novos dados.

Este trabalho utilizou o aprendizado supervisionado, que é a abordagem na qual um modelo preditivo é treinado com um conjunto de dados rotulados, ou seja, um conjunto de exemplos em que tanto os dados de entradas quanto as saídas desejadas (rótulos ou respostas desejadas) são fornecidas. O objetivo do aprendizado supervisionado é que o modelo aprenda a mapear as entradas para as saídas com base nos exemplos fornecidos, de modo que ele possa fazer previsões ou classificações precisas para novos dados não vistos anteriormente.

Nesse contexto, veremos a seguir algumas das técnicas de aprendizado de máquina amplamente utilizadas em cenários supervisionados em tarefas de classificação. Assim, a subseção seguinte apresenta os fundamentos das Redes Neurais Artificiais e Redes Neurais Convolucionais de modo a estabelecer o encadeamento teórico necessário para as discussões posteriores.

2.4.2 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) possuem ampla aplicação, abrangendo áreas como análise de imagens, processamento de sinais, controle, reconhecimento de padrões e diversas outras aplicações que demandam a capacidade de aprendizado a partir de dados de entrada. Trata-se de uma técnica de aprendizado de máquina que se inspira no funcionamento dos neurônios biológicos para processar dados de entrada. A sua origem em sua forma mais básica remonta a 1943, quando o neurofisiologista Warren McCulloch e o matemático Walter Pitts propuseram um modelo de neurônio artificial composto por circuitos elétricos (HAYKIN, 2001).

Ainda de acordo com o autor, podemos organizar os neurônios em camadas. Em sua configuração mais fundamental, temos uma camada de entrada de informações, na qual os nós abrigam as fontes de dados e se projetam na direção da camada de saída, que contém nós computacionais (ou seja, neurônios). Além disso, destacamos as redes

de múltiplas camadas, nas quais uma ou mais camadas ocultas são necessárias entre a camada de entrada e a camada de saída.

2.4.3 Redes Neurais Convolucionais

Dentre os algoritmos existentes para o reconhecimento e classificação de imagens, podemos destacar as CNNs. Esta é uma Rede Neural Profunda, devido à quantidade de camadas ocultas, e se baseia na percepção visual humana, sendo capaz de reconhecer formas bidimensionais com alto grau de invariância em relação a diversas formas de distorção (HAYKIN, 2001).

As CNNs são uma classe de algoritmos de DL que utilizam em sua arquitetura camadas convolucionais que aplicam filtros sobre os dados de entrada, permitindo a extração de características relevantes enquanto preservam as relações espaciais entre os elementos da imagem. Em problemas de classificação, as CNNs são capazes de aprender padrões e características distintas, geralmente resultando em um desempenho superior em comparação com métodos tradicionais de extração de atributos (TAJBAKHSH et al., 2016).

Dentre as CNNs, existem diversos modelos de arquiteturas, cada um com suas particularidades, que incluem tipicamente duas seções. A primeira corresponde a operações de convoluções e pooling, enquanto a segunda é composta de camadas totalmente conectadas, semelhantes às redes neurais convencionais. A Figura 7 mostra uma arquitetura genérica inspirada na LeNet-5 (LECUN et al., 1998).

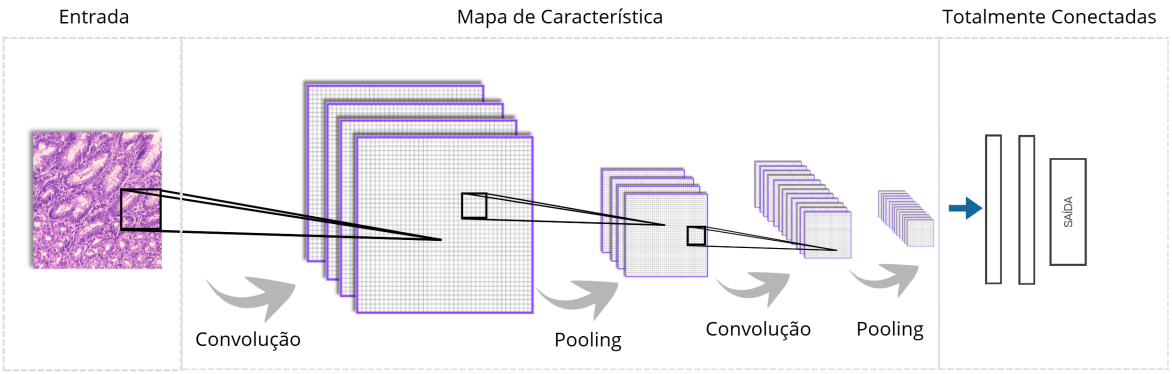


Figura 7 – Arquitetura genérica de CNNs. Fonte: Inspirado em (LECUN et al., 1998)

2.4.3.1 VGG

As arquiteturas VGGNet, propostas por Simonyan e Zisserman (SIMONYAN; ZISSERMAN, 2014), formam uma família de CNNs que se diferenciam pelo aumento sistemático da profundidade, isto é, pela variação do número de camadas convolucionais e totalmente conectadas. Originalmente, foram apresentadas seis variantes, contendo de 11

a 19 camadas treináveis, dentre as quais VGG16 e a VGG19 se tornaram as mais utilizadas.

A principal contribuição proposta pelas redes VGGNet foi o aumento da profundidade por meio do uso sistemático de filtros 3x3 em todas as camadas de convolução e operações de *max-pooling* 2x2 (janela 2x2, *stride* 2), que reduzem a largura e a altura dos mapas de características pela metade. A Figura 8 ilustra a arquitetura da VGG16.

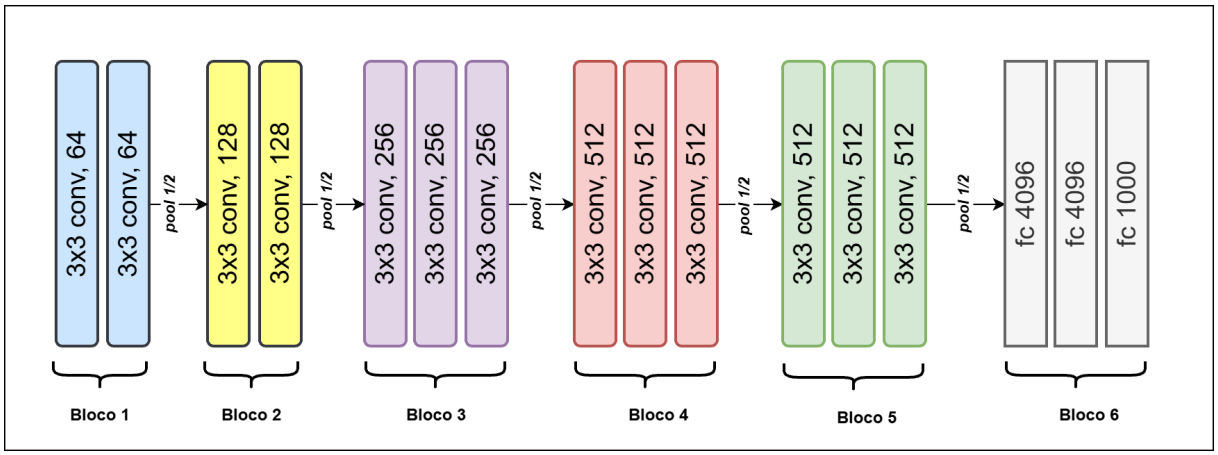


Figura 8 – Arquitetura da VGGNet com 16 camadas

2.4.3.2 Resnet

As CNNs profundas tornaram-se a melhor solução para obter avanços no processo de classificação de imagens, entretanto, o aumento da profundidade expõe o problema do desvanecimento do gradiente, no qual os gradientes utilizados para atualizar os pesos da rede tornam-se muito pequenos durante o treinamento. Para contornar esse desafio, a família ResNet introduziu a ideia de possibilitar de conexões de atalho viabilizando o desenvolvimento de CNNs com profundidade maior do que as tradicionais (HE et al., 2015).

O aprendizado residual, ilustrado na Figura 9, evidencia como a ResNet usa conexões residuais para permitir a propagação direta dos mapas de características de um ponto a outro da rede, sem que passem por todas as camadas convolucionais. As conexões residuais, ao preservar a transmissão eficiente da informação entre as camadas, facilitam a otimização dos parâmetros de redes mais profundas, reduzindo dificuldades associadas ao aumento da profundidade e evitando perdas significativas de desempenho da rede.

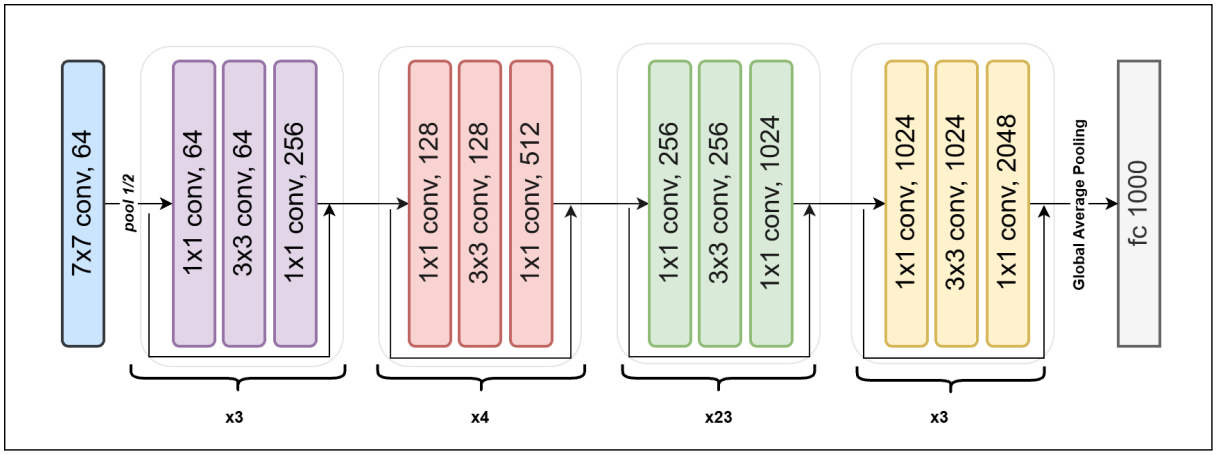


Figura 9 – Arquitetura da rede ResNet101, composta por uma camada convolucional inicial, seguida por blocos residuais e finalizada por *global average pooling* e uma camada totalmente conectada.

2.4.3.3 DenseNet

De forma semelhante à ResNet, Huang, Liu e Weinberger (HUANG et al., 2017) propuseram a DenseNet, uma arquitetura que tinha como objetivo mitigar o problema do gradiente desvanecente. Diferentemente da ResNet, que estabelece fluxos alternativos entre blocos de camadas, a DenseNet emprega conectividade densa, na qual cada camada recebe como entrada os mapas de características de todas as camadas anteriores, e suas saídas são concatenadas e encaminhadas às camadas subsequentes. Na Figura 10, mostra-se como os blocos densos agrupam várias camadas que concatenam suas saídas para serem passadas às camadas subsequentes.

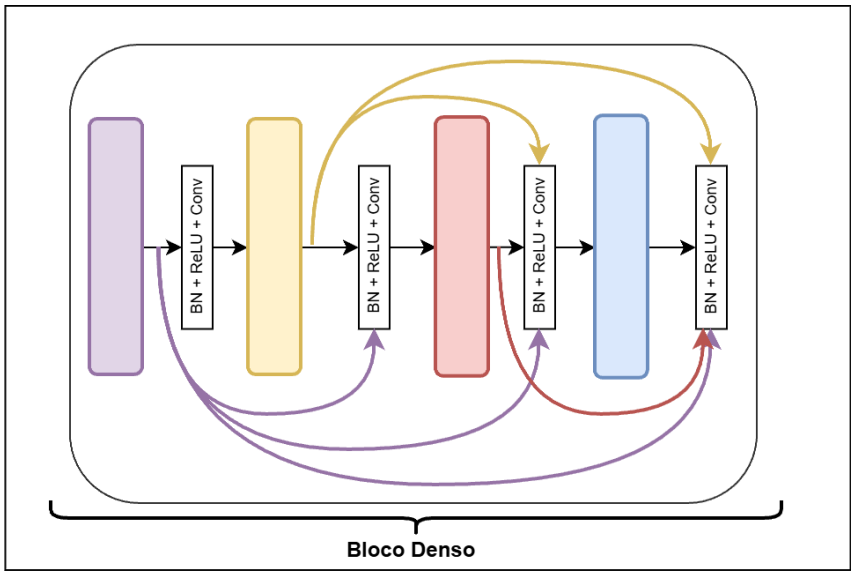


Figura 10 – Esquema de um bloco denso da DenseNet, em que cada camada recebe como entrada a concatenação das saídas de todas as camadas anteriores.

2.4.3.4 MobileNet

A cada nova CNN, o ganho de desempenho foi buscado com redes cada vez mais profundas e complexas, porém, esse caminho não tornavam necessariamente as redes mais eficientes em termos de tamanho e velocidade (SIMONYAN; ZISSERMAN, 2014). Nesse contexto, Howard *et al.* (HOWARD et al., 2017) propuseram as *MobileNets*, com uma arquitetura simplificada que usa convoluções separáveis em profundidade (*depthwise separable convolution*), que são uma forma de convoluções fatoradas, em que a convolução é dividida em uma convolução *depthwise* por canal e outra convolução 1x1, chamada de convolução pontual (*pointwise*). Essa decomposição reduz significativamente o custo computacional e o número de parâmetros, resultando em modelos substancialmente mais leves.

Uma convolução convencional realiza, em um único passo, a aplicação de um filtro e a combinação das entradas para gerar uma saída totalmente nova, como ilustrado na Figura 11-a. Nas convoluções separáveis em profundidade,o processo é dividido em uma camada separada para a filtragem (Figura 11-b) e outra separada para a combinação (Figura 11-c), reduzindo drasticamente a complexidade computacional e o tamanho do modelo (HOWARD et al., 2017).

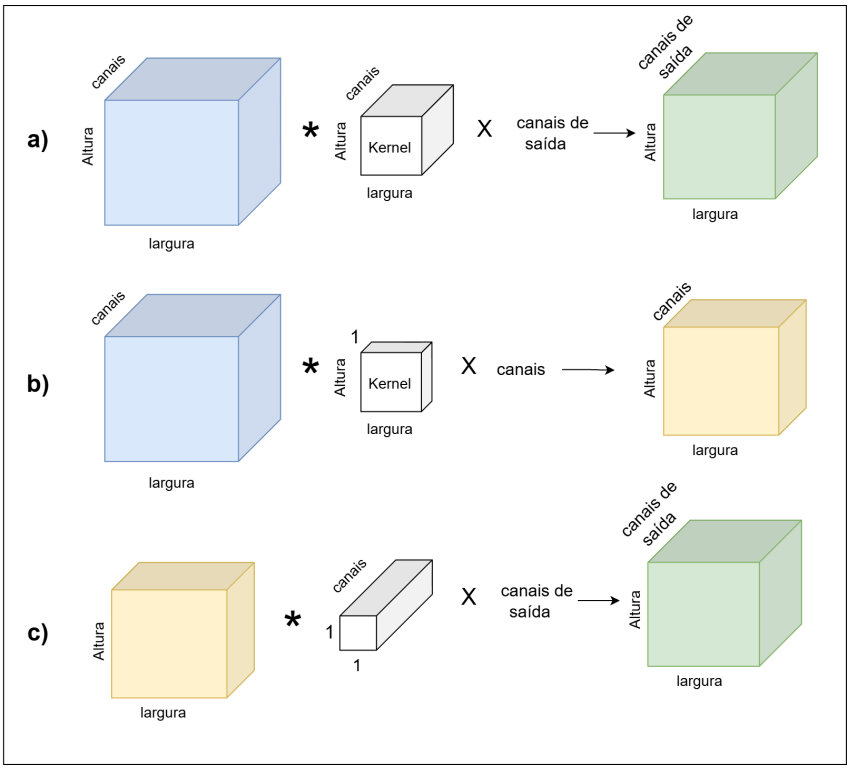


Figura 11 – Esquema comparativo entre (a) convolução 2D padrão, (b) convolução *depthwise* e (c) convolução 1×1 (*pointwise*), destacando a relação entre volume de entrada, filtros e volume de saída em cada caso.

2.4.3.5 *Efficientnet*

Tan e Le (TAN; LE, 2019) observaram que, para se obter melhor desempenho, outros modelos utilizavam a proposta de aumento da arquitetura, seja na largura (quantidade de canais), na profundidade (quantidade de camadas) ou na resolução das imagens. Isso deixava as redes mais pesadas, necessitando, assim, de um poder computacional maior. Desse modo, os pesquisadores apresentaram um processo mais simples de dimensionamento escalonado da rede, surgindo assim a família de arquiteturas da *EfficientNet*, que se baseia na ideia de equilibrar as dimensões de largura, profundidade e resolução para cada uma de suas redes. A Figura 12 apresenta a arquitetura da *EfficientNet_B0*.

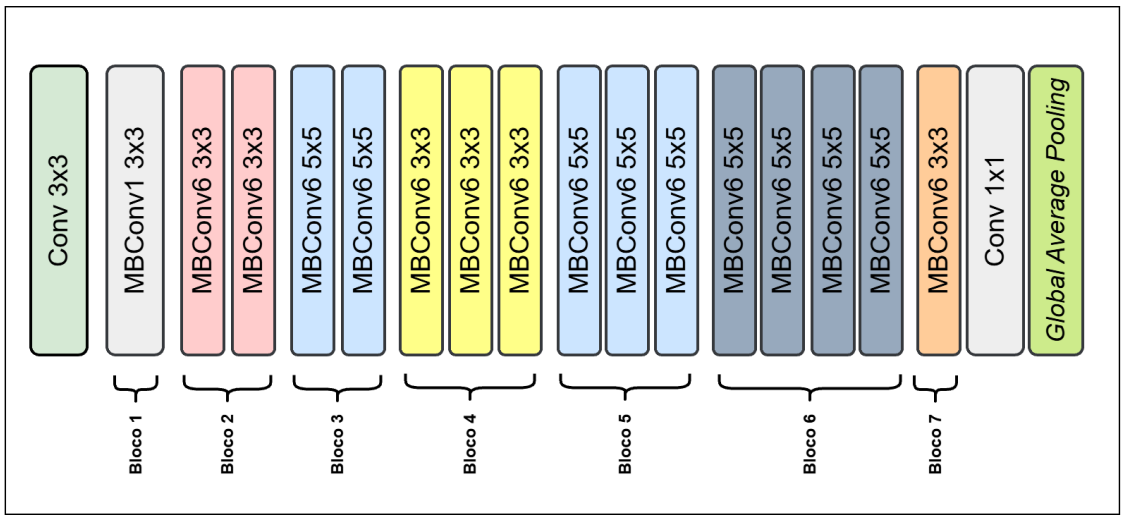


Figura 12 – Arquitetura da rede EfficientNet_B0 mostrando apenas as camadas convolucionais e o *global average pooling*.

2.4.3.6 *Vision Transformer*

Outro modelo de DL amplamente utilizado é o *Transformer*, responsável por grandes avanços no processamento de linguagem natural. Essa arquitetura possibilita explorar o poder de grandes quantidades de dados para obter uma compreensão mais profunda das informações subjacentes, conseguindo entender o contexto de uma palavra a partir das demais palavras da sequencia de entrada.

A sua arquitetura não usa convoluções, mas sim mecanismos de atenção, permitindo que o modelo foque dinamicamente nas partes mais importantes de uma sequência. Dessa forma, é possível para o modelo aprender relações complexas entre diferentes partes dos dados, considerando todas as posições da sequência de uma só vez (VASWANI et al., 2017).

Motivados pelo sucesso em processamento de linguagem natural, Dosovitskiy *et al.* (DOSOVITSKIY et al., 2021) propuseram o *Vision Transformer* (ViT)„ aplicando um *Transformer* padrão diretamente a imagens, com o mínimo possível de modificações.

Enquanto no texto o *Transformer* recebe como entrada uma sequência unidimensional, no caso de imagens a entrada 2D é particionada em uma grade de *patches*, com cada *patch* sendo achatado em um único vetor e submetido ao codificador *Transformer*. A Figura 13 ilustra a arquitetura do ViT.

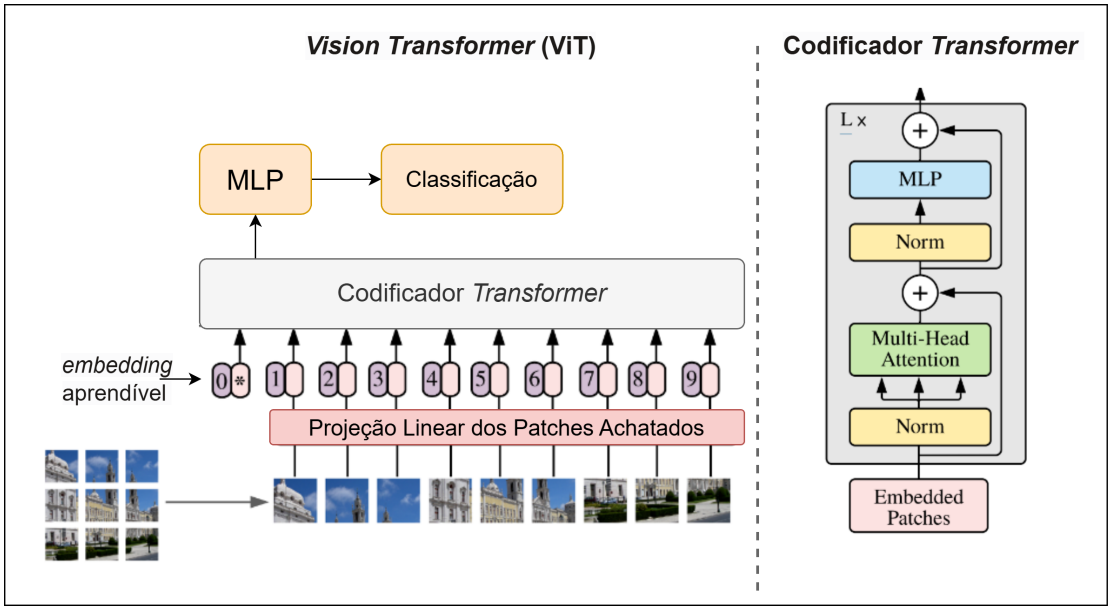


Figura 13 – Arquitetura geral do Vision Transformer (ViT), da divisão da imagem em *patches* ao codificador *Transformer* e à MLP de classificação. Fonte: Inspirado em Dosovitskiy et al. (2021).

2.5 Aumento de Dados

Algoritmos de DL têm se mostrado ferramentas capazes de produzir resultados superiores aos métodos tradicionais de ML, No entanto, a construção de modelos com alta capacidade de generalização tem sido uma tarefa desafiadora. Essa dificuldade ocorre devido às limitações dos conjuntos de dados, que nem sempre são grandes e heterogêneos. O problema da escassez de dados é mais evidente em *datasets* de imagens médicas, podendo ser atenuado por meio de métodos de ampliação de imagens (GOCERI, 2023).

O aumento de dados (*data augmentation*) é uma técnica de regularização, pois busca introduzir informações adicionais ao modelo com o objetivo de capturar melhor as propriedades gerais do problema em questão. Esse método tem um papel crucial na melhoria do desempenho e da capacidade de generalização dos modelos, alterando algumas características do conjunto de dados de treinamento (MUMUNI; MUMUNI, 2022).

O aumento de dados é comumente uma técnica utilizada para ampliar o *dataset*, seja com objetivo de balanceamento, seja para aumentar a variabilidade dos dados, porém, para contornar restrições relacionadas a armazenamento limitado, tem-se adotado o aumento de dados *in-place*, que transforma os dados em tempo real durante o trei-

namento, mesmo que isso deixe o processo mais lento (SHORTEN; KHOSHGOFTAAR, 2019; HAMEED et al., 2022).

O aumento de dados é comumente empregado para ampliar o *dataset* seja para balanceamento, seja para aumentar a variabilidade ; contudo, a fim de contornar restrições de armazenamento, tem-se adotado o aumento *in-place*, que transforma os dados em tempo real durante o treinamento, ainda que isso torne o processo mais lento (SHORTEN; KHOSHGOFTAAR, 2019; HAMEED et al., 2022).

2.6 Comitê de Classificadores

Um comitê de classificadores (*ensemble*) é uma estratégia bastante comum em problemas de classificação baseados em algoritmos de ML. Sua ideia central é combinar uma conjunto finito de modelos preditivos (classificadões base) para classificar novas amostras (Figura 14), melhorando a acurácia e a capacidade de generalização em relação aos classificadores individuais (DIETTERICH, 2000). Segundo Zhou (ZHOU, 2012), *ensembles* podem ser: (i) homogêneos, quando um único algoritmo gera múltiplos classificadores do mesmo tipo, treinados com conjuntos de dados ou hiperparâmetros levemente distintos; e (ii) heterogêneos, quando combinam-se algoritmos diferentes.

A eficácia de um *ensemble* está ligada a dois fatores principais, a diversidade entre os classificadores e a estratégia de combinação das predições individuais, de modo que o modelo em *ensemble* seja melhor do que qualquer um dos modelos individuais (GUEHRIA; BELLEILI; AZIZI, 2023). Dentre as estratégias de combinação, podemos destacar a média simples, média ponderada (por desempenho), votação majoritária e até a estratégia de *stacking* (empilhamento), neste último, as predições dos modelos base formam um conjunto de treino para um meta-modelo, que aprende pesos e interações, decidindo quando e em quem confiar.

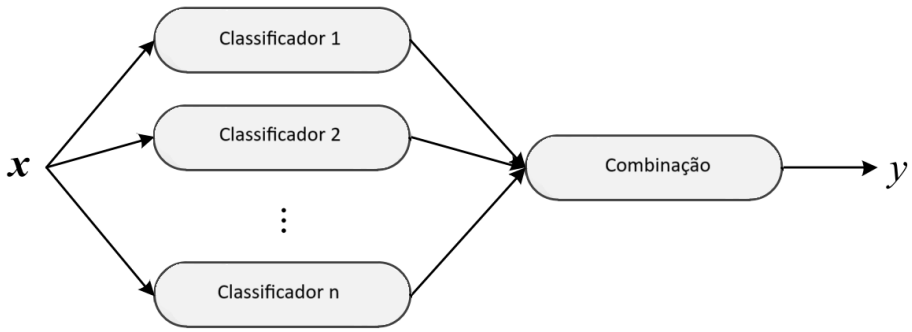


Figura 14 – Ilustração de uma arquitetura comum de *ensemble*. Fonte: Inspirado em (ZHOU, 2012)

2.7 Métricas de avaliação

Para avaliar o desempenho de modelos de DL, é essencial utilizar métricas de avaliação que garantam a representação adequada da eficácia e confiabilidade dos modelos analisados. Em problemas de classificação multiclasse, a matriz de confusão organiza, simultaneamente, os rótulos reais e as predições do classificador para todas as classes.

Para a tarefa de classificação de tecidos do TME, os conceitos de verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN) são analisados por classe, adotando a estratégia um-versus-todos (*one-vs-all*). A partir desses valores, são derivadas métricas como acurácia, precisão, *recall* e *F1-score* para cada classe, bem como suas médias agregadas para obtenção de indicadores globais de desempenho. Para cada classe específica, tem-se:

- VP: correspondem às amostras de uma classe que o modelo classificou corretamente como essa classe;
- VN: englobam todas as amostras que não pertencem a uma classe específica e que também não foram preditas como ela.
- FP: são amostras de outras classes que o modelo incorretamente classificou como a classe específica;
- FN: são amostras que pertencem a uma classe, mas foram preditas como outra;

A acurácia é uma métrica que calcula a proporção de todos os acertos de um classificador em relação ao total de amostras avaliadas. Em cenário multiclasse, a Equação 2.1 expressa a razão entre a soma dos verdadeiros positivos por classe (VP_c) e o número total de amostras avaliadas (N), sendo K o numero de classes possíveis. Neste estudo, a acurácia não será a principal métrica a ser analisada, pois, apesar de refletir o desempenho global do modelo, ela pode ocultar desempenhos insatisfatórios em classes específicas.

$$\text{acurácia} = \frac{\sum_{c=0}^{K-1} VP_c}{N} \tag{2.1}$$

Diferente da acurácia, a precisão é uma métrica que precisa olhar para cada classe específica em relação a todas as outras (*one-vs-all*). Ela quantifica o quanto o modelo é preciso quanto às predições positivas daquela classe; ou seja, dentre as amostras que o modelo classificou como uma classe específica, quanto realmente pertence a essa classe. A

Equação 2.2 define a precisão por classe, e a Equação 2.3 mostra que, para encontrar a precisão global, basta calcular a média de todas as precisões por classe.

$$\text{precisão}_c = \frac{VP_c}{VP_c + FP_c} \tag{2.2}$$

$$\text{precisão}_{global} = \frac{1}{K}\sum_{c=0}^{K-1} precisão_c \tag{2.3}$$

Assim como a precisão, o *recall* é uma métrica que olha para as classes; porém, aqui, ela busca mensurar quanto o classificador é sensível para uma determinada classe, ou seja, dentre todas as amostras de uma determinada classe, quantas o modelo conseguiu acertar. A Equação 2.4 e a Equação 2.5 mostram, respectivamente, como se calcula o *recall* por classe e o *recall* global.

$$recall_c = \frac{VP_c}{VP_c + FN_c} \tag{2.4}$$

$$recall_{global} = \frac{1}{K}\sum_{c=0}^{K-1} recall_c \tag{2.5}$$

Neste estudo, as métricas de precisão e *recall* são igualmente importantes, uma vez que a tarefa é a correta classificação dos tecidos do TME. Por essa razão, usaremos o *F1-score* como métrica de maior relevância em nossas avaliações, por sintetizar o equilíbrio entre precisão e *recall*. Matematicamente, o *F1-score* corresponde à média harmônica dessas duas medidas (Equação 2.6).

$$\textit{F1-score} = 2 \cdot \frac{\text{precisão} \cdot recall}{\text{precisão} + recall} \tag{2.6}$$

Também avaliamos a capacidade dos modelos de distinguir uma classe específica do conjunto das demais (*one-vs-all*), tratando esse cenário como um problema binário (positivo versus negativo). Para isso, utilizamos a Área sob a Curva (*Area Under the Curve* AUC), para mostrar quão bem um modelo separa amostras de uma classe de interesse daquelas que não pertencem a ela, criando assim uma AUC para cada classe do modelo.

A AUC é obtida a partir da Curva ROC (*Receiver Operating Characteristic*), que relaciona, ao longo de diferentes limiares, a taxa de verdadeiros positivos (*True Positive Rate* TPR), ou seja, o *recall*, e a taxa de falsos positivos (*False Positive Rate* FPR),

definida na Equação 2.7. Ao variar o limiar, calcula-se o par (TPR, FPR) e um gráfico é construído, sendo a área sob essa curva correspondente ao valor da AUC (Figura 15).

$$\text{FPR} = \frac{FP_c}{FP_c + VN_c} \tag{2.7}$$

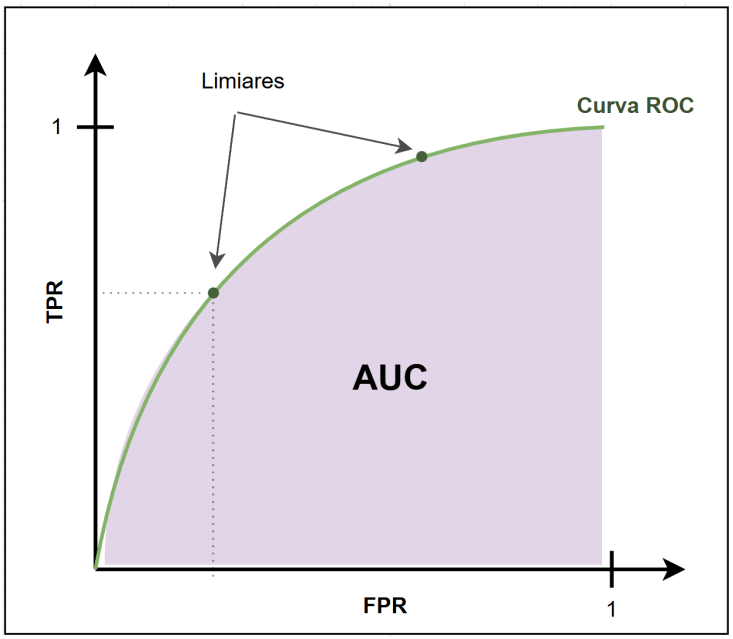


Figura 15 – Exemplo de curva ROC, destacando diferentes limiares de decisão e a área sob a curva (AUC).

2.8 Considerações Finais

Este capítulo consolidou a base teórica necessária para a compreensão do presente trabalho. Do ponto de vista médico, evidenciou-se a complexidade morfológica dos tecidos do TME e a demanda por ferramentas que auxiliem sua interpretação. Sob a perspectiva computacional, métodos contemporâneos de DL ofereceram uma alternativa promissora para a classificação de imagens histológicas, ainda que persistam desafios de generalização e de disponibilidade de dados anotados em CG. Assim, as reflexões aqui delineadas sustentam a formulação da metodologia proposta neste trabalho.

3 Trabalhos Relacionados

Este capítulo está organizado em quatro partes. Inicialmente, apresenta-se a estratégia de busca e os critérios de seleção dos estudos. Em seguida, são discutidos os trabalhos selecionados e apresentada uma síntese comparativa. Posteriormente, este trabalho é posicionado em relação à literatura revisada, destacando seus diferenciais. Por fim, são apresentadas as considerações finais do capítulo.

3.1 Estratégia de busca e seleção dos estudos

Este capítulo foi construído a partir de uma busca estruturada nas bases PubMed, Scopus, Web of Science e IEEE Xplore, utilizando termos como aprendizado profundo, CNN, histopatologia digital e microambiente tumoral (Figura 16). As publicações foram limitadas ao inglês e ao intervalo dos últimos cinco anos, ainda assim, publicações fora desse escopo foram encontradas em citações de trabalhos relacionados, como Kather *et al.* e Mikhailov *et al.* (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022), este último em idioma russo.

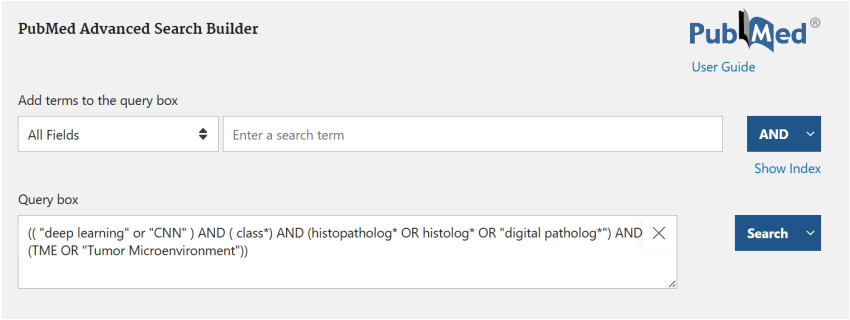


Figura 16 – Exemplo da *string* de busca utilizada na base da Pubmed.

Do conjunto inicial de estudos recuperados, foram incluídos neste capítulo apenas os trabalhos que: (i) utilizam algoritmos de *deep learning* aplicados a imagens histopatológicas, (ii) realizam alguma forma de classificação tecidual ou de núcleos celulares do TME e (iii) descrevem seus *datasets* de forma explícita. Os trabalhos apresentados a seguir representam exemplos dessas abordagens de seleção, com diferentes aplicações, não sendo limitados a um órgão específico.

3.2 Estudos selecionados

Na literatura existente, muitos autores se dedicaram a explorar o uso de algoritmos de IA para classificar, segmentar ou contabilizar componentes do TME, para as mais

diversas finalidades, desde propor um novo biomarcador até a predição de um evento de risco, como o óbito do paciente. No entanto, a classificação precisa desses componentes é essencial para garantir que as informações serão extraídas dos componentes corretos.

Na criação de biomarcadores, Kather *et al.* (KATHER et al., 2019) propuseram uma abordagem para quantificar diferentes componentes teciduais dentro do TME do câncer colorretal (CRC), desenvolvendo um escore chamado *Deep Stroma Score* para predizer a sobrevida desses pacientes. Os autores utilizaram imagens coradas com H&E de quatro conjuntos de dados diferentes: dois para a classificação dos componentes do TME (com nove classes) e dois para o cálculo do *Deep Stroma Score*. O processo de treinamento utilizou o *dataset* NCT-CRC-HE-100K, do biobanco NCT, com 100.000 recortes de imagens extraídas de 86 lâminas. A base de dados CRC-VAL-HE-7K, com 7.180 imagens, foi utilizada para validação externa do classificador. A CNN treinada foi utilizada para decomposição tecidual das bases do *The Cancer Genome Atlas* (TCGA), com 862 lâminas de 500 pacientes com dados clínicos completos, e do *Darmkrebs: Chancen der Verhütung durch Screening* (DACHS), com 409 lâminas de 409 pacientes. O modelo proposto pelos autores atingiu uma acurácia global de 94,0% na etapa de classificação.

Wang *et al.* (WANG et al., 2020) propuseram uma ferramenta capaz de segmentar e classificar núcleos celulares em imagens de adenocarcinomas do pulmão, com o objetivo de quantificar as células e correlacioná-las com a sobrevida dos pacientes. Foram utilizados dois conjuntos de dados: o *National Lung Screening Trial* (NLST), com 208 imagens, e o *The Cancer Genome Atlas Lung Adenocarcinoma* (TCGA-LUAD), com 431 imagens, ambos corados com H&E e com dimensão de 500 x 500 *pixels*. Após a segmentação dos núcleos celulares, eles foram classificados em seis classes de componentes do TME. A classificação foi feita de forma supervisionada, com as imagens rotuladas do *dataset* NLST, obtendo-se uma acurácia geral de 85,0%. Com relação ao modelo proposto para prognóstico de sobrevida, o TCGA-LUAD foi utilizado como *dataset* para validação externa.

No trabalho de *Rong et al.* (RONG et al., 2022), os autores propuseram um sistema voltado para a segmentação e classificação de núcleos celulares do TME. O estudo utilizou quatro conjuntos de dados distintos: o conjunto NLST de adenocarcinoma pulmonar, com 127 *patches* de 500×500 *pixels*, divididos em 6 tipos de núcleos; um *dataset* hepático, com 76 lâminas contendo tecidos humanos e de camundongos com ou sem carcinoma, classificados em 4 classes; o conjunto NuCLS, com 1.744 *patches* de câncer de mama oriundos do TCGA, divididos em 4 classes; e, por fim, o TCGA-BRCA, utilizado para análise de sobrevivência, contendo 1.061 lâminas com dados clínicos associados. Na tarefa de classificação dos núcleos em suas classes, a base de imagens do fígado teve acurácia de 84,4% sem apresentar os resultados de *F1-score*. Os resultados referentes a base NLST atingiram acurácia de 71,1% e *F1-score* de 74,1% . Finalmente a base com melhor

desempenho foi a NuCLS, que usou *ensemble* de modelos e conseguiu 86,6% de acurácia e 89,4% de *F1-score*.

No contexto da patologia digital aplicada ao câncer gástrico, o estudo de Mikhailov *et al.* (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022) propõe uma metodologia de anotação e rotulagem de imagens histológicas com foco no reconhecimento automático de tecidos da parede gástrica e na determinação da profundidade de invasão do câncer gástrico. A abordagem considerou cinco classes, visando otimizar o desempenho de algoritmos de classificação. O conjunto de dados utilizado foi o PATH-DT-MSU, desenvolvido no departamento de patologia do Centro Médico-Científico da Universidade Estatal de Moscou, que compreende 20 WSI anotadas manualmente a partir de fragmentos de parede gástrica. A extração de *patches* foi realizada por meio de uma janela deslizante, resultando em 70.871 imagens para treinamento e 14.462 para teste. Como classificador, foi utilizada uma rede DenseNet modificada, cujo desempenho evidenciou uma acurácia global de 91,7% e uma elevada acurácia na detecção da classe muscular da mucosa, considerada um marcador-chave para avaliação da invasão tumoral.

O trabalho de Zhou *et al.* (ZHOU et al., 2024) propõe uma assinatura de risco tumoral desenvolvida a partir de imagens histopatológicas de pacientes com câncer colorretal em estágio III, utilizando técnicas de aprendizado profundo. Para isso, empregaram um modelo de classificação tecidual baseado na arquitetura VGG19, capaz de distinguir nove tipos distintos de tecidos do TME, com uma acurácia global de 94,4%. O modelo foi treinado com o *dataset* NCT-CRC-HE-100K, disponibilizadas por Kather *et al.* (KATHER et al., 2019) e foram normalizadas usando o método *macenko*. As representações extraídas foram utilizadas para treinar modelos de risco com o objetivo de predizer a sobrevida do paciente.

No estudo apresentado por Wang *et al.* (WANG et al., 2024), os autores propuseram um conjunto de dados com 12 tipos diferentes de colorações de imunoistoquímica (IRQ) aplicados ao câncer do pulmão, permitindo a aplicação de modelos de DL para classificação de componentes do TME. O *dataset* foi composto a partir de 47 lâminas, coletadas do *Liaoning Cancer Hospital & Institute*, que foram anotadas manualmente por dois patologistas, os quais identificaram 7 tipos de tecidos. A fragmentação em regiões de 128×128 *pixels* foi seguida por sobreposição, descartando-se todos os *patches* que tivessem menos de 50% de alguma classe de tecido, chegando assim a 309.698 imagens. Nos resultados de classificação, os modelos VGG16 e o *Swin Transformer_tiny* apresentaram os melhores desempenhos, atingindo acurácia geral e *F1-score* de 81,1%.

Um estudo semelhante ao anterior foi o de Lou *et al.* (LOU et al., 2025), em que os autores desenvolveram um extenso conjunto de dados de imagens histológicas do TME de CG, denominado HMU-GC-HE-30K. Trata-se do conjunto de dados adotado no presente trabalho e, até onde se tem conhecimento, do único estudo que o utiliza. As imagens foram

obtidas a partir de 300 lâminas coradas com H&E, provenientes do *Cancer Hospital of Harbin Medical University*. As lâminas foram divididas em 31.096 *patches* de 224 × 224 *pixels*, e utilizou-se o classificador proposto por Kather *et al.* (KATHER et al., 2019) para dividi-las em oito classes de tecidos. Para a classificação, os autores empregaram os modelos ViT e EfficientNet-B0, ambos com pesos pré-treinados. Um subconjunto correspondente a 20% dos dados foi separado para os testes, sendo o melhor resultado alcançado pelo EfficientNet-B0, alcançando 70% de acurácia global e 71,0% de *F1-score*. A Tabela 1 mostra uma síntese comparativa entre os estudos revisados e este trabalho, contemplando: campos como objetivo, tipo de câncer, tamanho do *dataset*, número de classes, se utiliza técnica de *ensemble*, presença de estratégias de explicabilidade das predições e *F1-score* reportado.

Tabela 1 – Síntese comparativa dos trabalhos relacionados.

Autor/Ano	**Objetivo**	**Tipo**	**Dataset**	**Classe**	**Ensemble**	**Explicabilidade**	***F1-score***
Kather et al. (2019)	Biomarcador Sobrevida Classificar	CRC	100K 7K	9	Não	Sim	-
Wang et al. (2020)	Segmentar Classificar Sobrevida	Pulmão	208 imagens, 431 imagens	6	Não	Sim	-
Rong et al. (2022)	Segmentar Classificar Sobrevida	Pulmão; Fígado; Mama	127 *patches* 76 lâminas 1.744 *patches*	6 4 4	Não Não Sim	Não	74,1% - 89,4%
Mikhailov, Khvostikov e Krylov (2022)	Classificar	CG	85k	5	Não	Não	-
Zhou et al. (2024)	Classificar Risco tumoral	CRC	100K	9	Não	Não	-
Wang et al. (2024)	Criar *dataset* Classificar	Pulmão	47 lâminas 309.698 *patches*	7	Não	Não	81,1%
Lou et al. (2025)	Criar *dataset*; Classificar	CG	31.096	8	Não	Não	71%
Este trabalho	Classificar	CG	31.096	8	Sim	Sim	81.1%

3.3 Posicionamento deste trabalho na literatura

O presente estudo distingue-se do conjunto de trabalhos relacionados por um conjunto de características, resumidas a seguir:

(a) **Foco em câncer gástrico**: ao contrário de boa parte dos estudos analisados, que se concentram em câncer do pulmão (WANG et al., 2020; RONG et al., 2022; WANG et al., 2024) ou câncer colorretal (KATHER et al., 2019; ZHOU et al., 2024), este trabalho aborda técnicas de DL em imagens do TME de câncer gástrico, um cenário ainda pouco explorado na literatura

(b) **Classificação multiclasse de componentes do TME**: assim como Kather *et al.* (KATHER et al., 2019) e Zhou *et al.* (ZHOU et al., 2024), este trabalho realiza a classificação de múltiplos componentes teciduais do TME, considerando número elevado de classes, com a diferença de que o *dataset* usado neste trabalho não faz uso da classe fundo (*background*).

(c) **Uso do conjunto HMU-GC-HE-30K**: o estudo utiliza o *dataset* HMU-GC-HE (31.096 imagens) proposto por Lou *et al.* (LOU et al., 2025), sendo o único trabalho, até onde se tem conhecimento, além dos autores originais, a explorar essa base de dados em profundidade.

(d) **Avaliação de diferentes arquiteturas de DL**: a exemplo de Wang *et al.* (WANG et al., 2024) e Rong *et al.* (RONG et al., 2022), este estudo compara o desempenho de diferentes modelos pré-treinados (12 redes), buscando identificar arquiteturas mais adequadas para às particularidades do TME gástrico.

(e) **Explicabilidade**: o trabalho emprega técnicas de explicabilidade baseadas em mapas de ativação (Grad-CAM) para analisar quais regiões dos *patches* histológicos são mais relevantes para o modelo, o que favorece a interpretação dos resultados por especialistas.

(f) **Ênfase em comitês**: diferentemente da maioria dos trabalhos analisados, este trabalho aborda o uso de comitês para criar um classificador combinando redes que se complementam, a fim de obter uma classificação mais precisa do que outros estudos que utilizam a mesma base de dados.

(g) **Futuros biomarcadores**: embora o foco principal seja a classificação tecidual do TME em CG, os resultados obtidos podem servir como base para a construção de classificadores de risco, como o *Deep Stroma Score* (KATHER et al., 2019), ao cruzar padrões teciduais associados a desfechos clínicos.

3.4 Considerações Finais

Neste capítulo, observamos que um aspecto comum entre os trabalhos analisados é a utilização de algoritmos de DL, em especial CNNs, aplicados a diferentes tarefas envolvendo o TME, como segmentação, criação de biomarcadores, predição de risco e criação de *datasets*. Apesar da diversidade de objetivos, todos os trabalhos precisaram realizar tarefas de classificação dos componentes do TME em alguma etapa do processo, reforçando o papel central dessa tarefa para a extração de informações clinicamente relevantes.

De modo geral, os resultados de acurácia dos trabalhos situam-se acima de 80%, sendo menos comuns valores superiores a 90%, e apenas parte dos trabalhos apresenta o

F1-score. Além disso, poucos estudos exploram explicitamente estratégias de *ensemble* ou técnicas de explicabilidade, e a maioria concentra-se em câncer de pulmão ou colorretal. Vale ressaltar que o conjunto de dados utilizado nesta pesquisa não foi utilizado em nenhum dos estudos analisados, à exceção de Lou *et al.* (LOU et al., 2025), que o propôs originalmente.

4 Materiais e Métodos

Neste capítulo, descreve-se o metodologia proposta, contemplando o *dataset* utilizado e os experimentos realizados para se chegar ao modelo final, cujo objetivo é classificar os diferentes tipos de tecido do TME do câncer gástrico. O fluxograma da metodologia adotada é representado na Figura 17, iniciando pela preparação dos dados, passando pelo treinamento dos modelos, até chegar a formação dos comitês.

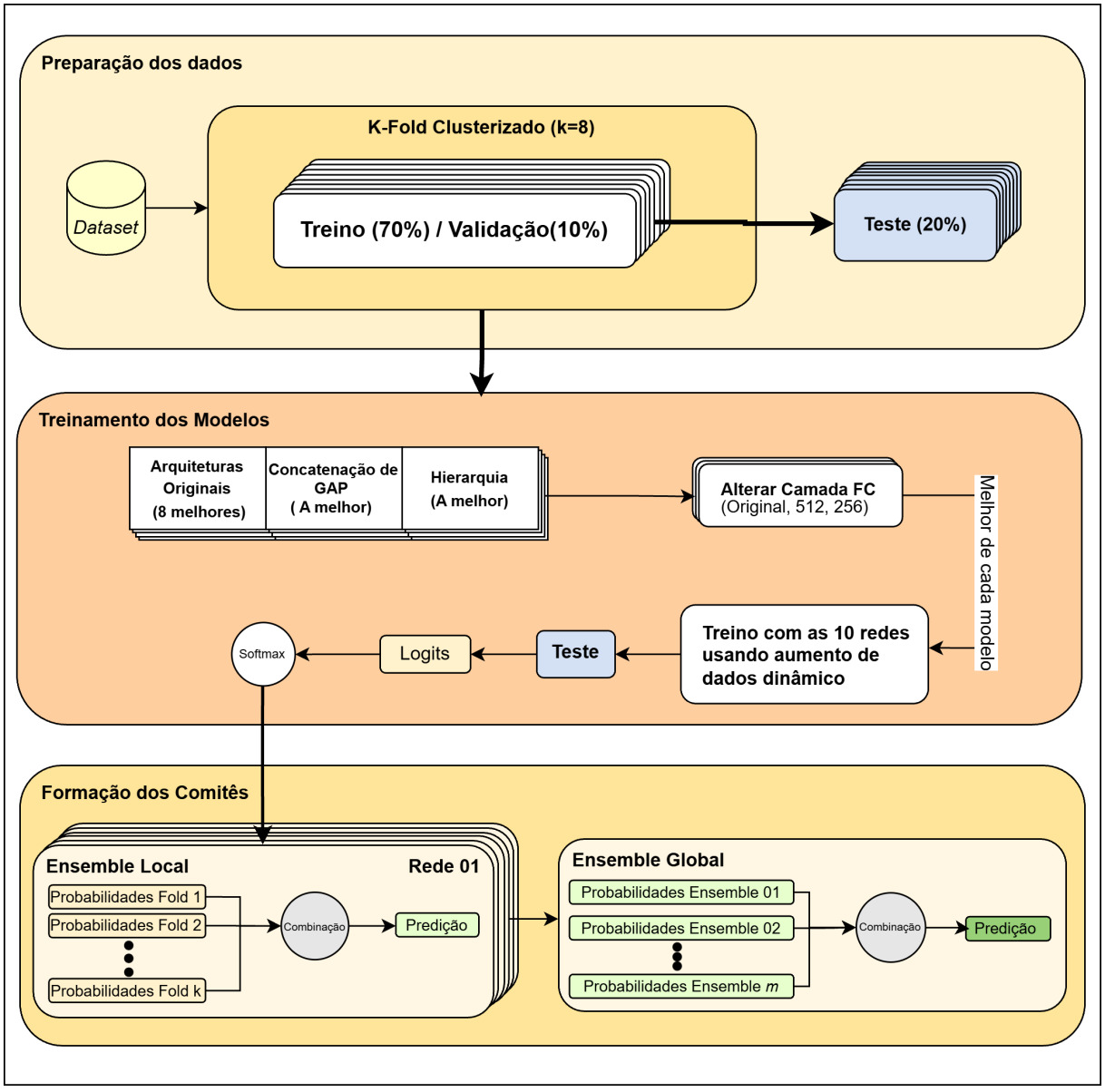


Figura 17 – Fluxo dos experimentos.

4.1 Preparação dos dados

Na primeira fase do estudo, foi realizada uma análise exploratória do *dataset*, com o objetivo de prepará-lo adequadamente para o processo de treinamento dos modelos. As imagens foram separadas de forma estratificada em treino, validação e teste, considerando os agrupamentos naturais decorrentes de diferenças na coloração e na iluminação das lâminas (variações de brilho e tonalidade). O intuito dessa abordagem visa garantir uma avaliação mais realista do desempenho dos modelos, respeitando a heterogeneidade dos dados.

4.1.1 Conjuntos de dados

A base de dados utilizada no presente trabalho foi a HMU-GC-HE-30K, apresentada por Lou *et al.* (LOU et al., 2025) em seu artigo. O *dataset* é composto por imagens obtidas a partir de 300 lâminas coradas com H&E, provenientes de diferentes pacientes do Hospital de Câncer da Universidade Médica de Harbin, no período de 2013 a 2015. Os dados foram aprovados pelo comitê de ética e disponibilizados publicamente na plataforma *Figshare* (LOU et al., 2024).

A aquisição das imagens histológicas foi realizada com o scanner Aperio AT2 (*Leica Biosystems*, Alemanha), equipado com lente de ampliação de 20 vezes. As WSI foram fracionadas em 31.096 imagens originais não sobrepostas, com *patches* de 224×224, divididos igualmente em oito classes: tecido adiposo (ADI), detritos (DEB), muco (MUC), músculo (MUS), agregados de linfócitos (LYM), estroma (STR), mucosa normal (NOR) e epitélio tumoral (TUM), como ilustrado na Figura 18.

O processo de anotação foi conduzido em três etapas: a primeira foi uma pré-anotação feita utilizando um modelo treinado com o conjunto de dados de Kather *et al.* (KATHER et al., 2019), com imagens de CRC; em seguida, uma etapa de verificação manual com dois patologistas juniores, ambos com mais de 5 anos de experiência na área; e, por fim, uma etapa de checagem final feita por um patologista sênior, com mais de 10 anos de experiência.

Após todo o processo de anotação do *dataset*, cada uma das oito classes ficou composta por 3.887 imagens, não havendo necessidade de balanceamento posterior. Para validar os dados, Lou *et al.* (LOU et al., 2025) utilizaram uma arquitetura baseada em Transformer (ViT) e um modelo baseado em CNN (EfficientNet-B0) para a análise de classificação dos oito componentes teciduais do TME.

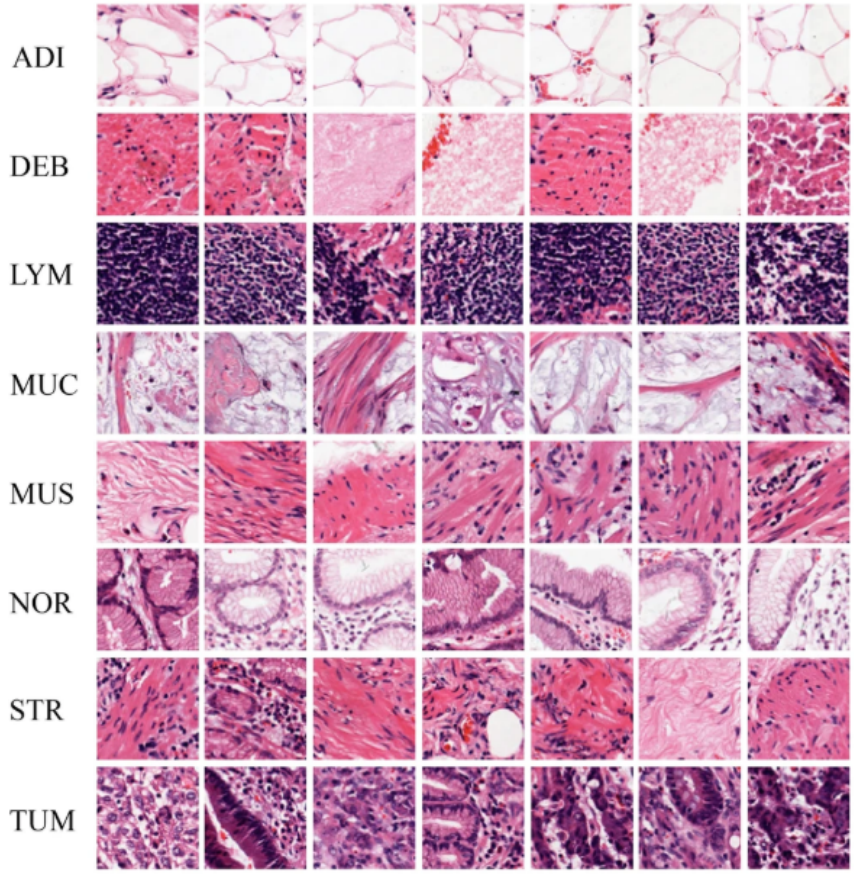


Figura 18 – *Patches* extraídos rotulados com as oito classes de tecido TME. Fonte: (LOU et al., 2025)

4.1.2 Validação Cruzada

Este trabalho utilizou a técnica de validação cruzada *k-fold* estratificada, na qual o conjunto de dados é distribuído de forma aleatória em k subconjuntos (*folds*), preservando a proporcionalidade das classes do conjunto de dados original. Com a quantidade de k execuções, haverá uma combinação diferente de treino e validação para cada uma delas. Essa técnica é amplamente utilizada para avaliar a capacidade de generalização de um modelo.

Nos experimentos realizados, optou-se por utilizar a distribuição das imagens em 70/10/20, por ser esta a divisão mais frequentemente reportada na literatura. Foram reservados 20% de todas as imagens do *dataset* como um conjunto de testes independente, ou seja, não serão utilizadas no treinamento. O remanescente foi submetido à validação cruzada e, para assegurar a distribuição global de 70% para treino e 10% para validação, adotou-se $k = 8$, como é ilustrado na Figura 19.

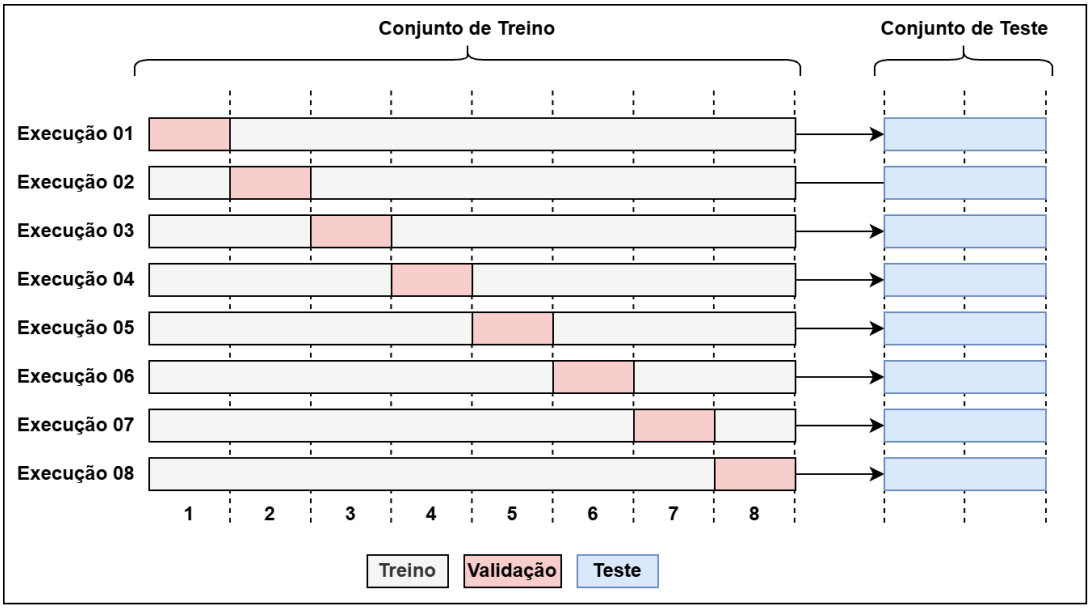


Figura 19 – Distribuição de validação cruzada com k=8 e 20% do total dos dados separados para teste.

Devido às imagens serem provenientes de diferentes WSIs, elas apresentam variação de brilho e cor dentro de uma mesma classe. Deste modo, para garantir que a heterogeneidade intra-classe fosse mantida em todos os *k folds*, inspirados no trabalho de Liang *et al.* (LIANG et al., 2024), aplicamos separadamente em cada classe do conjunto de treinamento, o algoritmo *k-means* buscando agrupar os *patchs* em 4 *clusters*, de forma que cada grupo represente um padrão intra-classe distinto.

Em seguida, cada *fold* é construído de forma estratificada por classe e por *cluster*, onde as amostras são aleatoriamente e divididas, garantindo que, em cada *fold*, todas as classes estejam presentes e que, dentro de cada classe, haja amostras provenientes dos seus quatro *clusters*. Esse procedimento evita que, por acaso, algum *fold* fique composto apenas por imagens muito semelhantes (por exemplo, de um único padrão intra-classe), o que poderia comprometer a estabilidade dos resultados. A Figura 20 ilustra esse processo.

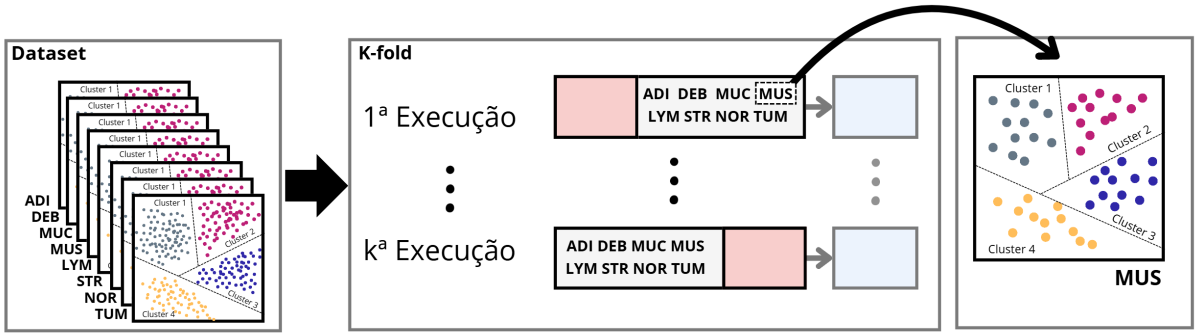


Figura 20 – Distribuição das imagens no conjunto de treino utilizando agrupamento intra-classe. À esquerda, o *dataset* é dividido em *k folds*. À direita, exemplo da classe MUS: cada ponto representa um *patch* de imagem, colorido de acordo com o cluster ao qual pertence.

4.2 Modelagem dos Modelos

Considerando a limitação de dados em *datasets* de imagens do TME de CG, foi adotada a estratégia de transferência de aprendizado (*Transfer Learning* TL), permitindo que iniciássemos o treinamento com algum conhecimento prévio sobre padrões visuais. Essa abordagem consiste em adaptar um modelo pré-treinado em uma grande base de dados para uma nova tarefa, reutilizando conhecimento previamente adquirido. Neste estudo, foram utilizadas arquiteturas com pesos pré-treinados no conjunto de dados *ImageNet* (DENG et al., 2009).

Nesta etapa de modelagem das redes, foram exploradas diferentes abordagens para extrair o melhor desempenho dos modelos, para que, na etapa seguinte, pudéssemos construir os comitês. Os experimentos foram organizados em fases, e nesta seção, serão descritas as estratégias exploradas, enquanto os resultados serão apresentados no próximo capítulo.

4.2.1 Fase 1: Experimentos com Arquiteturas Originais

Inicialmente, optou-se por testar redes pré-treinadas recorrentes em estudos de classificação do TME com aprendizado profundo. Foram incluídas EfficientNet (TAN; LE, 2019), DenseNet121 e DenseNet201 (HUANG et al., 2017), MobileNetV2 (HOWARD et al., 2019), XceptionNet (CHOLLET, 2017), VGG16 (SIMONYAN; ZISSERMAN, 2014), ResNet18, ResNet50 e ResNet101 (HE et al., 2016), além da AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Tabela 2 – Lista das arquiteturas avaliadas.

Nome	**Abbr.**	**Parâmetros treináveis**
EfficientNet-B0	*efn*	4.040.000
EfficientNet-B2	*efn_b2*	7.712.266
DenseNet121	*dsn121*	7.218.312
DenseNet201	*dsn201*	18.108.296
ViT	*vit*	86.787.080
MobileNetV2	*mbn*	4.212.280
Xception	*xcp*	20.823.344
VGG16	*vgg*	138.357.544
ResNet18	*rsn18*	11.180.616
ResNet50	*rsn50*	23.500.000
ResNet101	*rsn101*	44.606.536
AlexNet	*alx*	57.036.616

4.2.2 Fase 2: Experimentos com Foco nos Vetores de Características

Na Fase 1 dos experimentos, observou-se que o conjunto de dados apresenta algumas classes com maior índice de conflito do que outras (ADIxMUC, MUSxSTR, e NORxTUM). Segundo Mikhailov *et al.* (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022), o estroma tumoral (STR) é particularmente difícil de rotular por se encontrar de forma difusa entre as células tumorais (TUM). Os autores também destacam que a mucosa normal e o tecido tumoral tendem a se confundir, dado o limiar sutil de separação entre essas duas estruturas.

Inspirados no trabalho de Hameed *et al.* (HAMEED et al., 2022), que propõe a extração de características de camadas convolucionais intermediárias e sua posterior concatenação como entrada para o classificador, desenvolvemos uma abordagem baseada na utilização dos vetores de características (*embeddings*) extraídos no processo de convolução da rede. O objetivo é adaptar nosso modelo para melhor distinguir as classes mais conflitantes.

Na nossa abordagem, optamos por modificar a arquitetura *efn* (21-a), por ser a mesma utilizada por Lou *et al.* (LOU et al., 2024) ao apresentar o *dataset*. Durante o processo convolucional, extraímos os *embeddings* por meio da técnica *Global Average Pooling* (GAP), que calcula a média de cada mapa de características, resultando em vetores representativos. Esse processo é realizado em quatro etapas distintas da convolução e, ao final, os vetores resultantes são concatenados e passados ao classificador. O fluxo completo desse processo está ilustrado na Figura 21-b.

Outras duas abordagens foram utilizadas seguindo essa mesma metodologia. Na primeira, representada na Figura 21-c, foi adicionado um novo GAP após a primeira camada convolucional, com o intuito de incluir características morfológicas na composição

final do vetor. Na segunda abordagem, visando simplificar o vetor concatenado, optou-se por utilizar apenas os dois últimos GAPs das abordagens anteriores, conforme ilustrado na Figura 21-d.

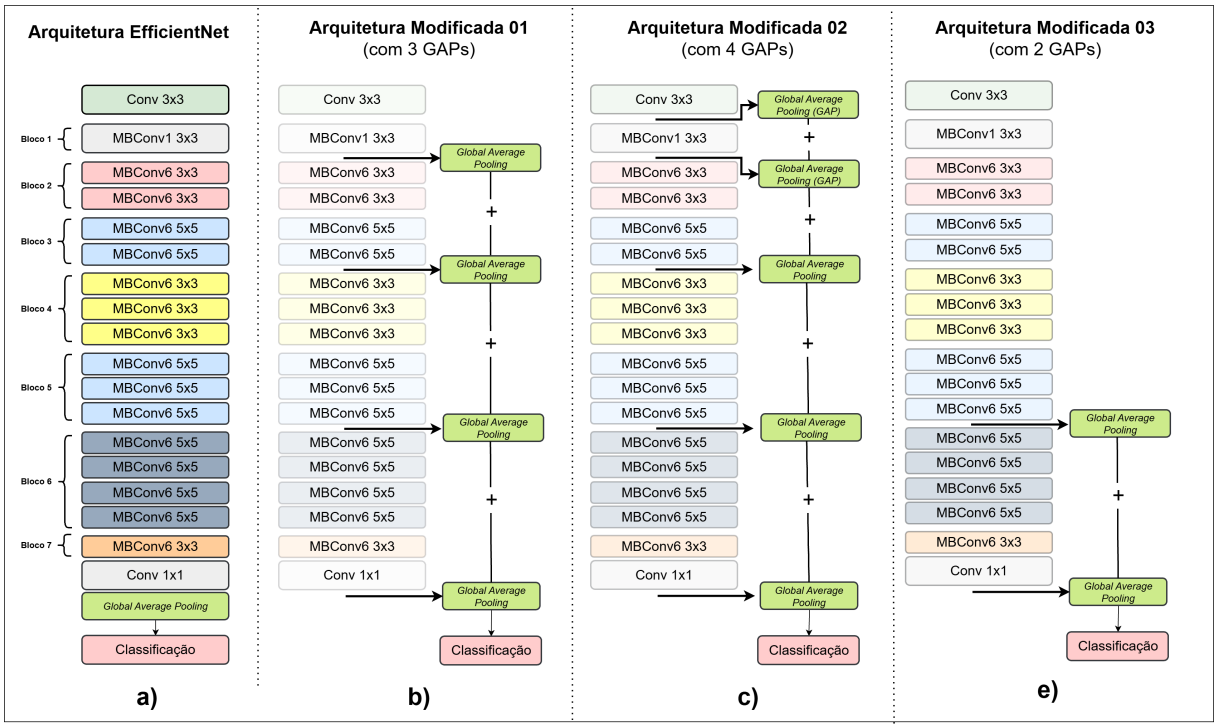


Figura 21 – Diferentes estratégias de extração de características utilizando operações de GAP. (a) Arquitetura Original;(b) Estratégia 4 GAPs; (c) inclusão de um GAP adicional após a primeira camada convolucional; (d) versão simplificada com apenas os dois últimos GAPs

4.2.3 Fase 3: Experimentos com Hierarquização de treinamentos

Nesta fase dos experimentos, na tentativa de solucionar os principais conflitos encontrados, utilizamos a metodologia hierárquica proposta por Zheng *et al.* (ZHENG et al., 2020), que estrutura o processo de treinamento e predição em níveis. Essa metodologia foi aplicada usando duas abordagens distintas: na primeira, foi treinado um modelo para classificação multiclasse considerando todas as classes da base, e adicionalmente, modelos de classificação binária para cada par de classes com maior conflito. Ao avaliar este cenário no conjunto de teste, sempre que uma predição envolve um desses pares de classes, o algoritmo redireciona a entrada para o classificador binário correspondente.

Na segunda abordagem, dividimos as classes em uma sequência de classificadores binários. A Figura 22 ilustra esse processo, no qual, primeiramente, agrupamos as classes em dois grupos: um com baixo conflito com o tecido tumoral (lado A) e outro de alto conflito com o tecido tumoral (lado B). Em seguida, hierarquizamos os pares dentro de cada lado, realizando diversos treinamentos binários.

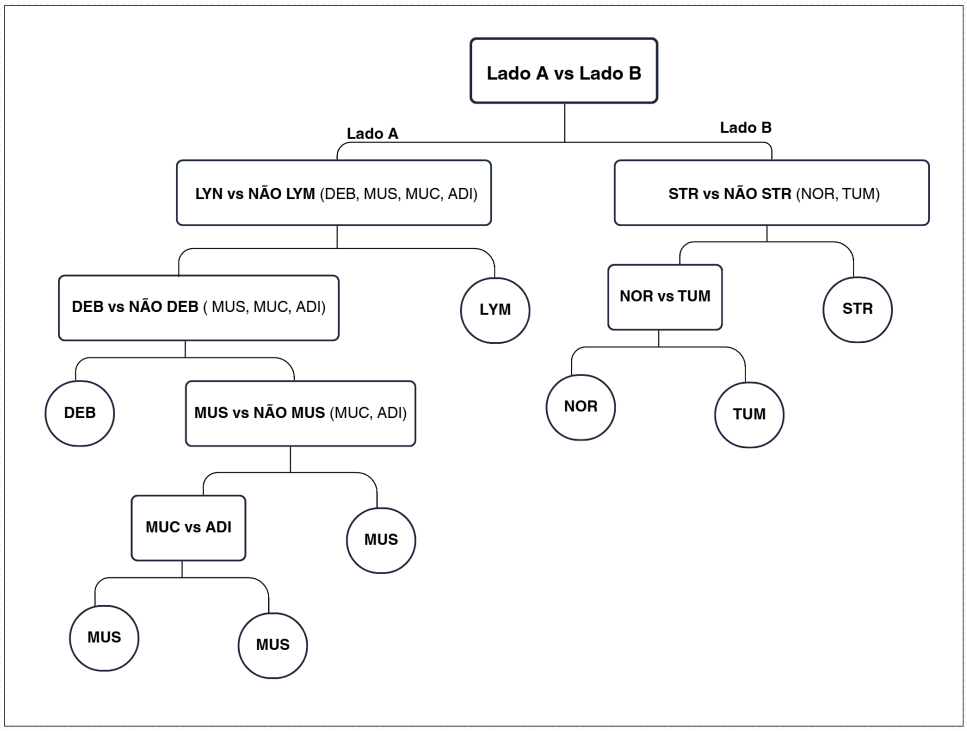


Figura 22 – Agrupamento hierárquico das classes, apresentando baixo nível de conflito (lado A) e alto nível de conflito (lado B) em relação ao tecido tumoral.

Durante o processo de predição, cada imagem é submetida à hierarquia de classificadores binários. No exemplo ilustrado na Figura 23, o processo se inicia no nível 1, que determina se a imagem pertence a um grupo com ou sem conflito com tecido tumoral. Em seguida, a imagem é direcionada ao lado A (sem conflito com tumor), sendo então submetida às avaliações subsequentes: no nível 2, verifica-se se é LYM ou não LYM; no nível 3, DEB ou não DEB; no nível 4, MUS ou não MUS; e, por fim, no nível 5, se é MUC ou ADI. Caso o exemplo correspondesse a uma imagem com conflito, o sistema decidiria primeiro entre STR e não STR e, caso não fosse STR, realizaria a classificação entre NOR e TUM.

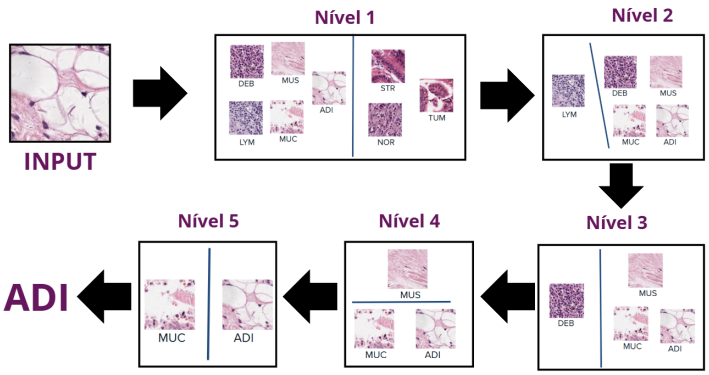


Figura 23 – Exemplo de fluxo de decisão hierárquica com classificadores binários, resultando na classe final.

4.2.4 Fase 4: Experimentos Focados em Quantidade de Neurônios

Na quarta fase dos experimentos, explorou-se uma abordagem inspirada no trabalho de Santos *et al.* (SANTOS et al., 2022), que investiga especificamente o impacto da capacidade da camada de classificação sobre o desempenho do modelo. Para esta fase, foram escolhidas as oito redes com os melhores *F1-score* da Fase 1, além da melhor da Fase 2 e da melhor da Fase 3, totalizando dez modelos. Então avaliamos se o limite de desempenho estava no extrator de características ou na camada totalmente conectada responsável pela decisão final.

Para responder a esse questionamento, modificamos apenas a camada totalmente conectada das redes, variando a quantidade de neurônios para 256 e 512 em cada uma das dez arquiteturas selecionadas, com o objetivo de verificar se modificações na camada densa melhora a separação entre as oito classes de tecido do TME. Na Figura 24, ilustra as alterações realizadas nessa etapa.

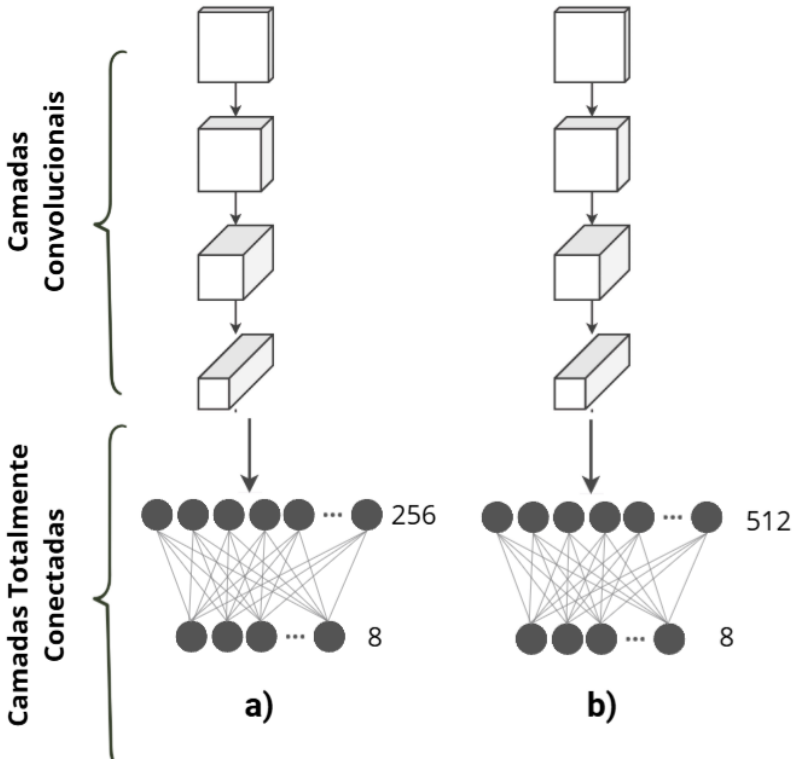


Figura 24 – Arquiteturas com modificações na camada totalmente conectada diminuindo a quantidade de neurônios (256 e 512).

4.2.5 Aumento de dados

O presente trabalho utilizou a biblioteca *Albumentations* (IGLOVIKOV et al., 2025), que aplica transformações durante o treinamento, em tempo de execução (*in-place*). Diferentemente do aumento *off-line*, que amplia o conjunto de dados armazenado ao criar novas imagens no disco, o aumento *in-place*, não aumenta a quantidade de amostras por

época, mas sim a diversidade observada pelo modelo, pois a cada época, cada amostra pode ser estocasticamente transformada (com uma probabilidade pré-definida), gerando amostras sintéticas. Assim, ao longo do treinamento, há um aumento cumulativo de variabilidade, sem modificar o balanço de classes nem o número de iterações por época.

As transformações foram realizadas com base em trabalhos relacionados e combinadas de maneira empírica até obtermos os melhores resultados. As técnicas utilizadas incluem: (i) **SquareSymmetry** (Figura 25-a), que aplica rotações simétricas (90°, 180°, 270°), espelhamentos (vertical e horizontal) e inversões (vertical, horizontal, diagonal e anti-diagonal); (ii) **Affine** (Figura 25-b), responsável por translações, rotações, *zoom* e cisalhamento; (iii) **RandomToneCurve** (Figura 25-c), que ajusta a intensidade de tom simulando variações de iluminação; e (iv) **RandomGridShuffle** (Figura 25-d), que divide a imagem em grades e embaralha os blocos, gerando uma nova composição visual. Cada uma dessas transformações possui a probabilidade de 30% de ser aplicada a cada imagem, de forma não exclusiva.

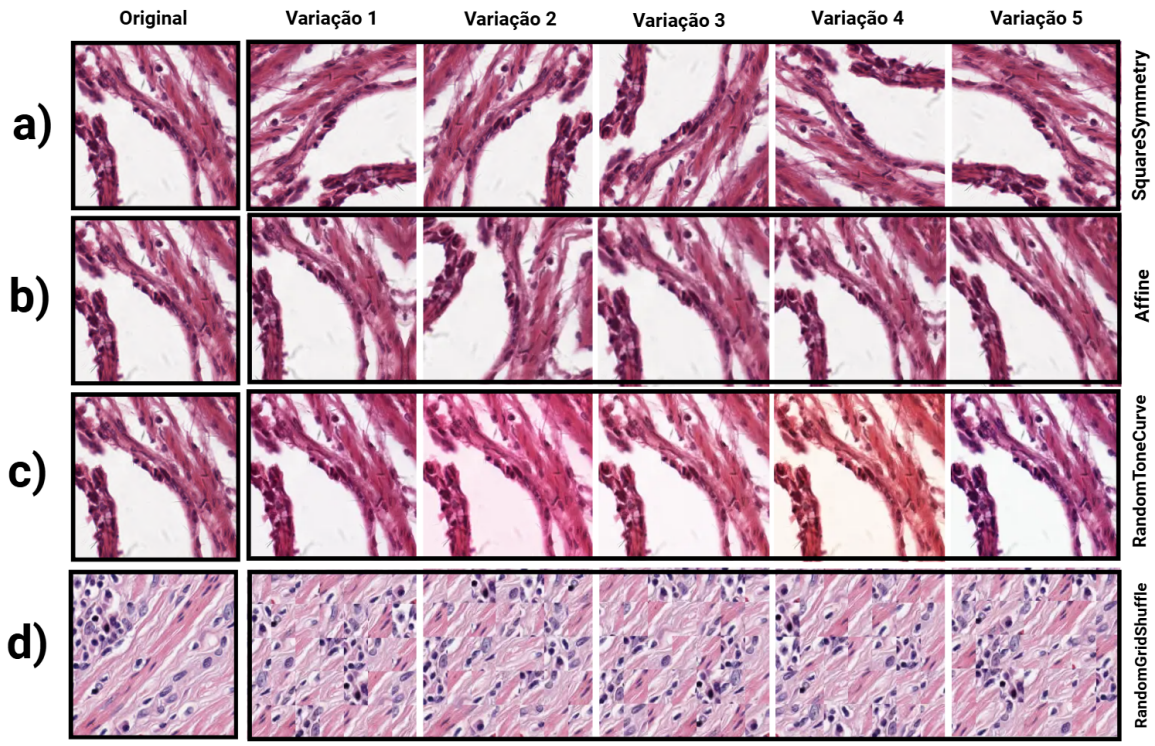


Figura 25 – Imagens dos aumentos utilizados. a) Imagem original seguidas de transformações de rotação, espelhamentos horizontal e vertical, inversão diagonal e anti-diagonal. b) Imagem original seguidas de translação, rotação, *zoom out, zoom in* e cisalhamento; c) Imagem original seguida variações de cor; d) imagem original seguida de composições diferentes de blocos embaralhados

4.3 Estratégias de treinamento

Durante a etapa de treinamento adotou-se a estrategia *On-Step Tuning* (OST), na qual todas as camadas da rede são ajustadas conjuntamente (treinamento fim a fim), mantendo a ordem convencional de propagação, em que as camadas convolucionais extraem características seguidas pelas camadas totalmente conectadas responsáveis pela classificação. O treinamento da rede foi realizado por várias épocas, utilizando uma condição de parada (*Early Stop*) configurada para interromper o processo após 10 épocas consecutivas sem melhora na função de perda (*loss*).

A Tabela 3 apresenta os detalhes da otimização de hiperparâmetros realizada por meio do *GridSearch*, utilizando o modelo *efn* como referência e aplicando os melhores resultados obtidos às demais redes. Os intervalos dos hiperparâmetros foram definidos com base na revisão da literatura e em estudos anteriores. Entre os hiperparâmetros avaliados estão: a taxa de aprendizado (*learning rate* - LR), os algoritmos otimizadores (Adam e SGD), o parâmetro de regularização *weight decay*, (*batch size*) e *dropout*. A busca pela melhor combinação foi conduzida por esta técnica, a qual explora exaustivamente as combinações possíveis dentro dos intervalos especificados.

Tabela 3 – Resumo dos hiperparâmetros testados.

Hiperparâmetros	**Range**	**Melhor Resultado**
LR	1×10^{-3}, 1×10^{-4}, 9×10^{-5}	9×10^{-5}
Otimizador	Adam, SGD	Adam
weight decay	1×10^{-3}, 1×10^{-4}	1×10^{-4}
Batch Size	32, 64, 128	64
Dropout	0.2, 0.3, 0.4, 0.5	0.5

O computador utilizado para execução dos experimentos conta com a seguinte configuração de *hardware*: processador Intel i5 de 12ª geração, 16GB de memória RAM e uma placa gráfica GeForce RTX 3060 com 12GB de memória dedicada. Todo o processo de treinamento dos modelos foi implementado com biblioteca PyTorch com Python na sua versão 3.11.7.

4.4 Formação de Comitês

Na construção dos *ensembles*, dois aspectos são fundamentais: a seleção dos membros e a estratégia de combinação. Propomos realizar essas tarefa em dois níveis hierárquicos: um *ensemble* local, no qual os membros são escolhidos entre classificadores da mesma arquitetura, e um *ensemble* global, composto por modelos com arquiteturas distintas.

4.4.1 Ensemble Local

Após a etapa de treinamento, cada modelo dispôs de 8 classificadores base, correspondentes a cada execução da validação cruzada, que serão os membros do *ensemble* local. A abordagem foi inspirada em Dutschmann *et al.* (DUTSCHMANN et al., 2023), que empregam validação cruzada repetida para gerar membros e combinar por média, e em Zhang e Liu (ZHANG; LIU, 2023), que utilizam esse procedimento para aprender pesos aplicados em etapa subsequente.

A Figura 26 ilustra como as imagens de teste são avaliadas por esses classificadores, produzindo *logits* que, em seguida, são transformados em probabilidades pela ativação *softmax*. Tais probabilidades são então agregadas por uma função de apuração para gerar uma predição final.

Embora todos os modelos do comitê compartilhem a mesma arquitetura e os mesmos hiperparâmetros, cada execução da validação cruzada é treinada com partições de dados distintas e com diferentes realizações das operações de aumento de dados, o que resulta em ajustes de pesos específicos para cada modelo. Essas diferenças tornam os classificadores parcialmente complementares, permitindo que o *ensemble* homogêneo reduza a variância das predições e mitigue o viés pessimista associado ao uso isolado de um único *fold*.

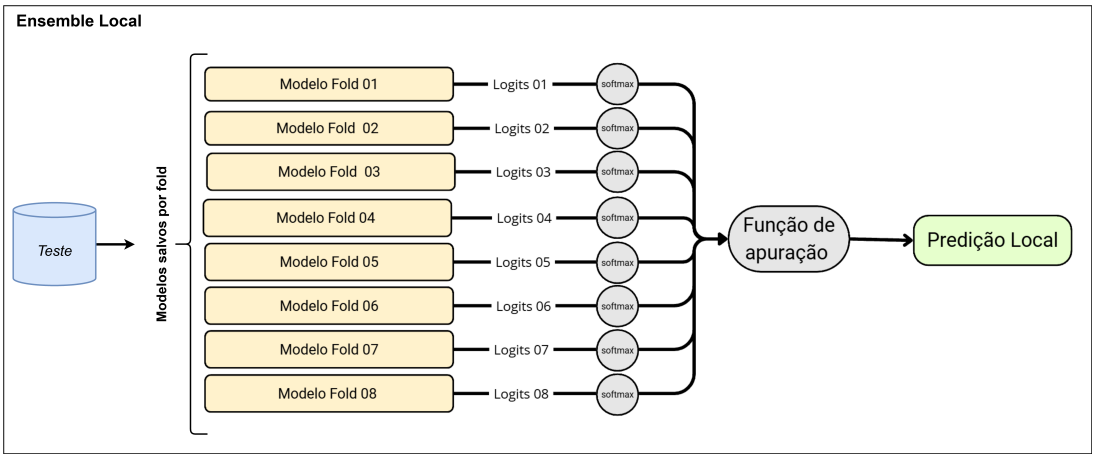


Figura 26 – Representação do *ensemble* local. Cada um dos oito classificadores base, derivados das diferentes execuções da validação cruzada, gera uma predição individual que são então combinadas, resultando na predição local do comitê.

4.4.2 Ensemble Global

Na etapa de *ensemble* global, utilizaram-se as probabilidades combinadas nos comitês locais para realizar uma nova combinação, porém agora entre arquiteturas distintas. Nesta etapa, o critério de seleção foi feito combinando todas as possibilidades distintas, com três, cinco e sete membros, dentre as dez arquiteturas selecionadas. Para este estudo,

optou-se por utilizar como critério de desempate as métricas globais na seguinte ordem: *F1-score*, AUC, acurácia e *recall*. A ilustração da Figura 27 mostra o diagrama de um *ensemble* de dois níveis.

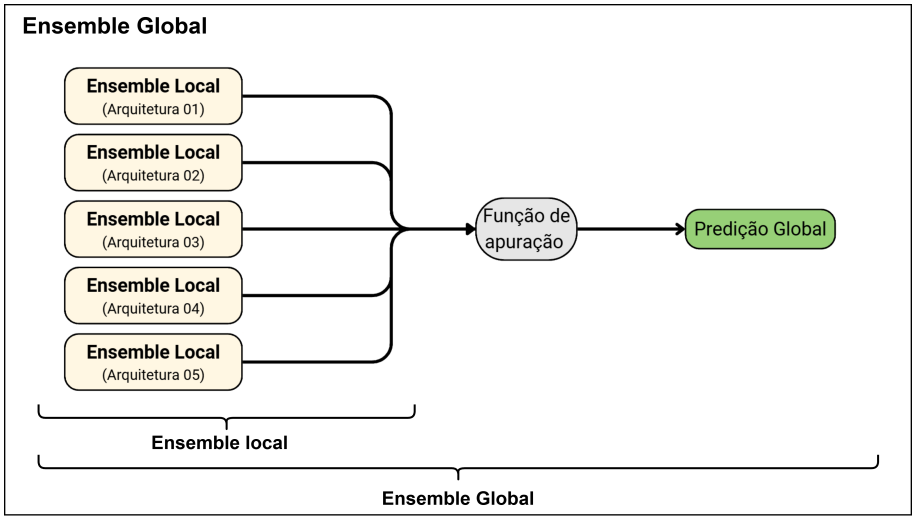


Figura 27 – Estrutura do *ensemble* global. As predições dos *ensembles* locais são combinadas para gerar a predição final.

4.4.3 Estratégia de Apuração

Uma das etapas fundamentais na construção de *ensembles* é a definição da estratégia de combinação, ou seja, em como as probabilidades serão apuradas para que uma nova predição seja realizada. A Tabela 4 apresenta nove estratégias organizadas em três grupos: Simples (sem ponderação), Modelo (ponderação pelo *F1-score* global de cada modelo) e Classe (ponderação pelo *F1-score* específico de cada classe). Em cada grupo, consideram-se três variantes: média, votação majoritária e média das probabilidades máximas de cada modelo, esta ultima inspirada no trabalho de Abnar e Makarenkov (ABDAR; MAKARENKOV, 2019) que usa votação majoritária mas leva em conta a confiança (probabilidade) do modelo.

Tabela 4 – Descrição dos métodos de combinação de classificadores utilizados.

Grupo	Estratégia	Descrição
Simples	**I**	Média aritmética das probabilidades dos modelos.
	II	Votação majoritária entre os modelos.
	III	Média apenas das probabilidades *Top-1* de cada modelo.
Modelo	**IV**	Combinação ponderada pelo F1-score global de cada modelo.
	V	Votação majoritária ponderada pelo desempenho global dos modelos.
	VI	Combinação ponderada apenas das probabilidades *Top-1*, por modelo).
Classe	**VII**	Combinação ponderada pelo F1-score por classe.
	VIII	Votação majoritária ponderada pelo desempenho por classe.
	IX	Combinação ponderada apenas das probabilidades *Top-1*, por classe).

A estratégia I, ilustrada na Figura 28, exemplifica a apuração realizada a partir das probabilidades de cada classe produzidas pelas predições dos diferentes membros do comitê. O procedimento consiste em calcular, para cada classe, a média aritmética (não ponderada) das probabilidades (apuração intraclasse), resultando em uma predição final formada pelas médias de cada classe, sendo a abordagem mais utilizada na literatura para fusão de decisões em comitês (GANAIE et al., 2022).

A Equação 4.1 mostra que $p(n, c)$ é probabilidade do *ensemble* para uma amostra n na classe c, que é dada pela média aritmética das probabilidades previstas pelos m membros do comitê para essa classe:

$$p(n, c) = \frac{1}{m} \sum_{i=0}^{m-1} p_i(n, c) \tag{4.1}$$

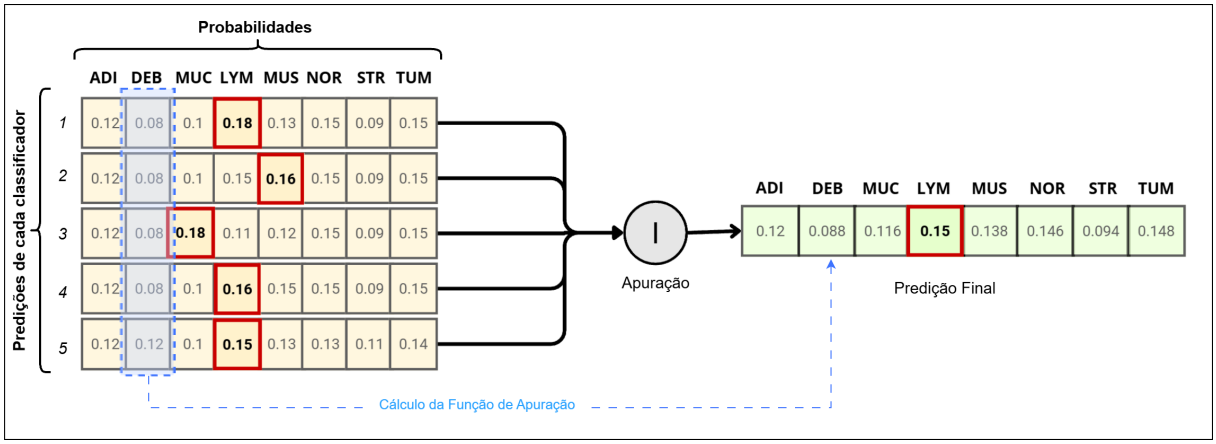


Figura 28 – Exemplo da Estratégia I de apuração: as probabilidades previstas pelos classificadores para cada classe são combinadas pela média aritmética, produzindo o vetor de probabilidades finais do *ensemble*.

A Estratégia II corresponde a uma votação majoritária, em que, para cada classificador, saída é binarizada: atribui-se 1 à classe com maior probabilidade e 0 às demais. Em seguida, os votos são combinados por classe, neste caso somando por rótulos (GANAIE et al., 2022), conforme a Equação 4.2, produzindo uma distribuição não normalizada. Para obter a predição final, aplica-se a normalização (Equação 4.4), em que cada valor é dividido pela soma total dos votos (Equação 4.3). Esta estratégia pode ser acompanhada na Figura 29.

$$p(n, c) = \sum_{i=0}^{m-1} p_i(n, c) \tag{4.2}$$

$$p_{\text{soma}}(n) = \sum_{i=0}^{c-1} p(n, c_i) \tag{4.3}$$

$$p_{\text{normalizado}}(n, c) = \frac{p(n, c)}{p_{\text{soma}}(n)} \tag{4.4}$$

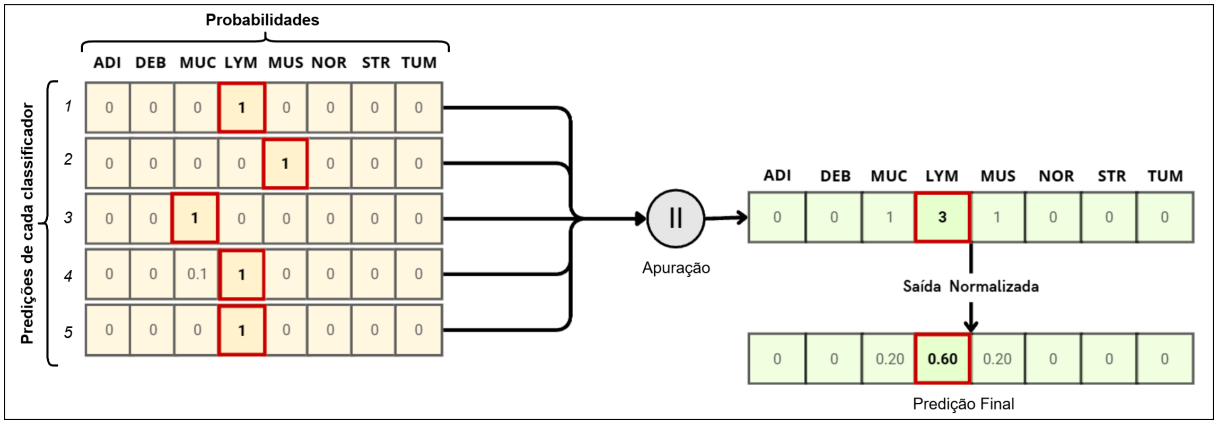


Figura 29 – Exemplo da Estratégia II de apuração: as predições discretas de cada classificador são somadas por classe (votação) e, em seguida, normalizadas para gerar o vetor de probabilidades finais do *ensemble*.

Na Estratégia III, as probabilidade associada à classe de saída, ou seja, a de maior valor são mantidas, anulando-se as demais (0), para cada classificador. Trata-se de uma variante da votação majoritária (Estratégia II) que preserva a intensidade da confiança do modelo na classe vencedora privilegiando predições com maior confiança e suprimir contribuições menos confiantes. A estratégia pode ser visualizada na Figura 30 e as equações que a definem são as 4.1, para média e 4.4 para a normalização.

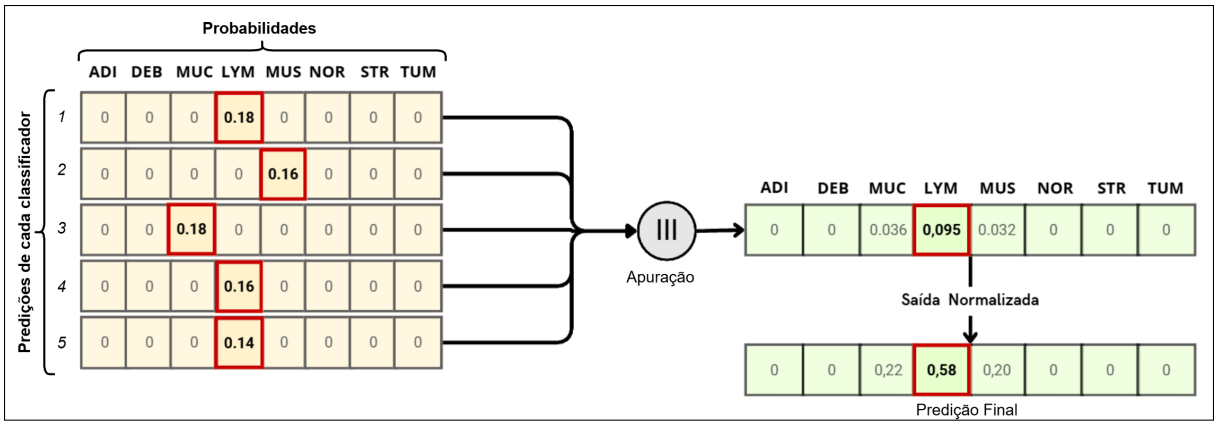


Figura 30 – Exemplo da Estratégia III de apuração: para cada classificador, apenas a maior probabilidade é mantida (demais tornam-se zero) e, em seguida, essas probabilidades são somadas e normalizadas para compor o vetor de probabilidades finais do *ensemble*.

Inicia-se o segundo grupo de estratégias, no qual as combinações utilizam ponderações baseadas nos valores normalizados de *F1-score* dos modelos que compõem o comitê.

Na Estratégia IV, a Equação 4.5 define que a probabilidade de cada classe na predição final é a soma ponderada das probabilidades, isto é, o somatório do produto entre o peso normalizado do membro e a probabilidade atribuída à classe por esse membro. Um exemplo dessa estratégia é apresentado na Figura 31.

$$p(n,c) = \sum_{i=0}^{m-1} w^i p_i(n,c) \tag{4.5}$$

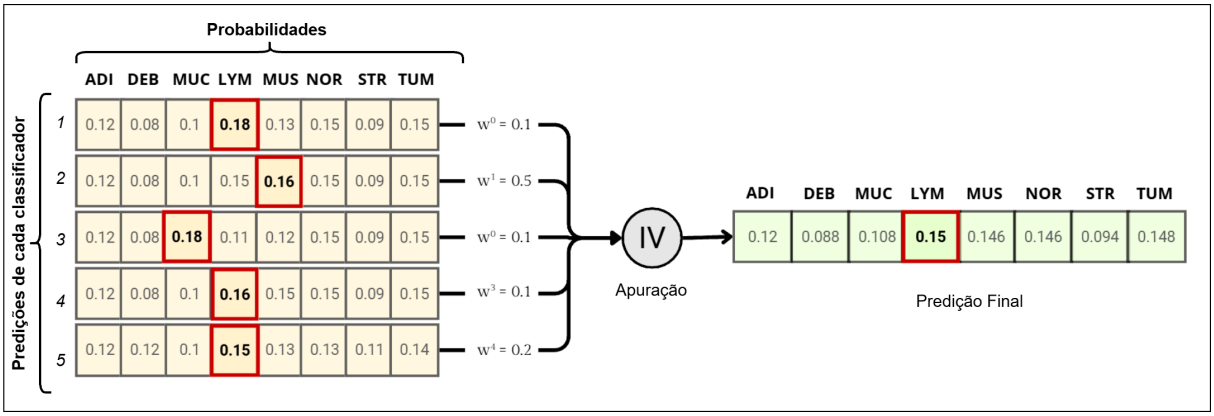


Figura 31 – Exemplo da Estratégia IV de apuração: as probabilidades previstas por cada classificador são combinadas por média ponderada, utilizando pesos normalizados baseados no *F1-score* de cada modelo.

A Estratégia V (Figura 32) e a VI (Figura 33) seguem o mesmo princípio das Estratégias II e III, respectivamente. Contudo, neste grupo não é necessária a normalização da predição final, pois os pesos derivados do *F1-score* já se encontram normalizados, assim, ao aplicar a Equação 4.5, obtêm-se probabilidades entre 0 e 1.

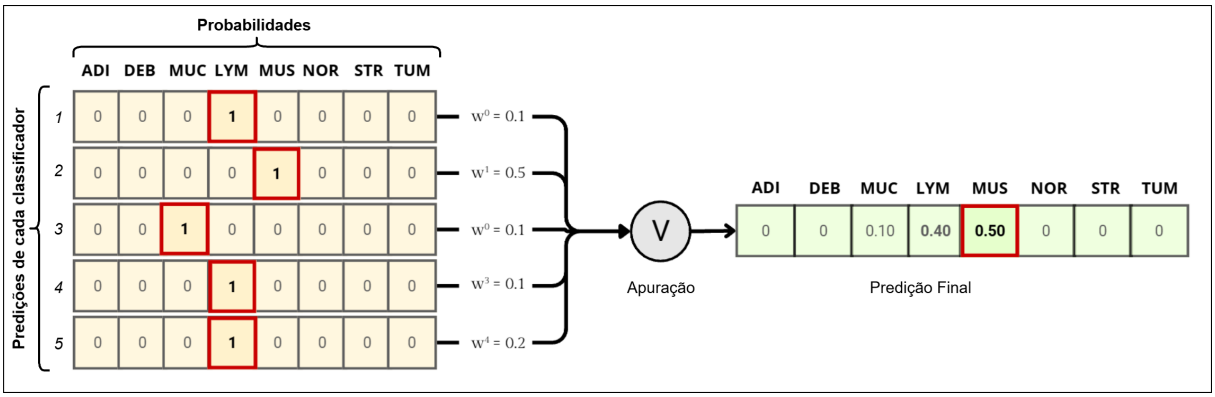


Figura 32 – Exemplo da Estratégia V de apuração: os votos discretos de cada classificador são ponderados pelos pesos normalizados de *F1-score* global e somados por classe, gerando as probabilidades finais do *ensemble*.

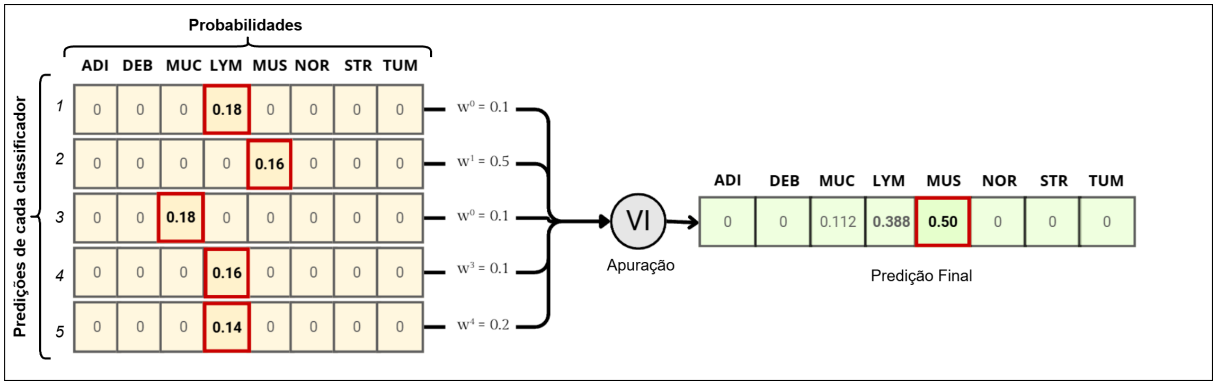


Figura 33 – Exemplo da Estratégia VI de apuração: para cada classificador, apenas a maior probabilidade é mantida e ponderada pelos pesos normalizados de *F1-score* global; em seguida, essas contribuições são somadas por classe para formar a predição final do *ensemble*.

No último grupo, as estratégias serão ponderadas pelo *F1-score* específico de cada classe. Assim, cada membro passa a possuir um vetor de oito pesos normalizados por classe. As Estratégias VII (Figura 34), VIII (Figura 35) e IX (Figura 36) utilizam a Equação 4.6 para calcular a soma ponderada por classe em seguida, aplicam e a Equação 4.4 para normalizar as probabilidades da predição final.

$$p(n, c) = \sum_{i=0}^{m-1} w_c^i \, p_i(n, c) \tag{4.6}$$

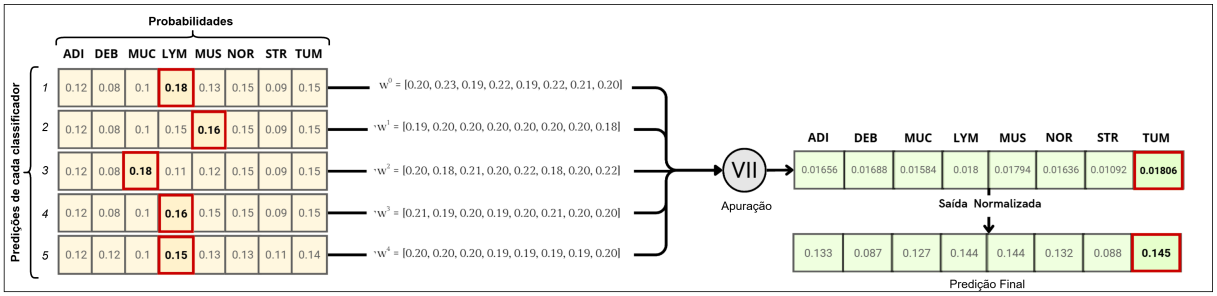


Figura 34 – Exemplo da Estratégia VII de apuração: as probabilidades previstas por cada classificador são ponderadas pelos pesos específicos de *F1-score* por classe e, em seguida, normalizadas para gerar o vetor de probabilidades finais do *ensemble*.

O objetivo desse grupo de apuração é privilegiar os modelos com melhor desempenho em classes específicas e atenuar a influência daqueles mais fracos em outras classes, reduzindo assim o risco de viés e potencializando a mitigação de confusão entre classes conflituosas.

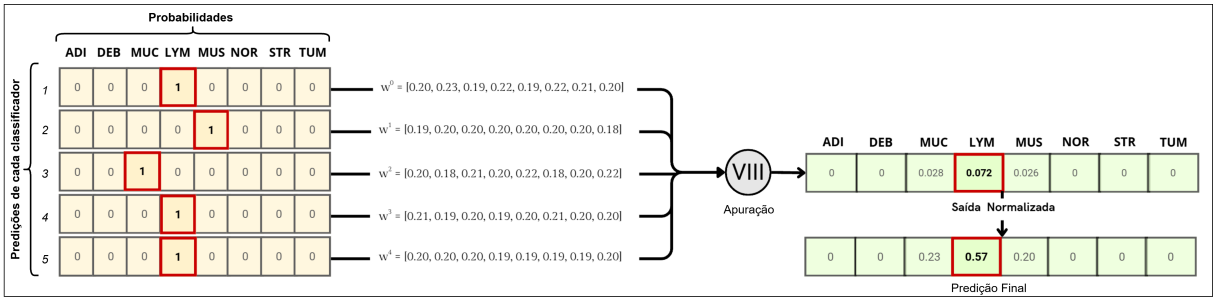


Figura 35 – Exemplo da Estratégia VIII de apuração: os votos discretos de cada classificador são ponderados pelos pesos específicos de *F1-score* por classe e, após normalização, formam o vetor de probabilidades finais do *ensemble*.

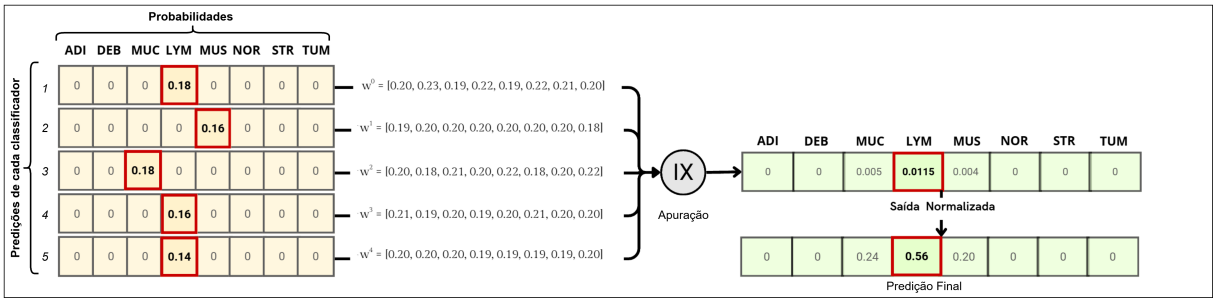


Figura 36 – Exemplo da Estratégia IX de apuração: para cada classificador, apenas a maior probabilidade é mantida, ponderada pelos pesos específicos de *F1-score* por classe e, em seguida, normalizada para compor o vetor de probabilidades finais do *ensemble*.

4.5 Reprodutibilidade

Esta pesquisa prezou pela reprodutibilidade em todas as etapas, desde a descrição metodológica detalhada até a disponibilização pública do código-fonte e dos artefatos necessários para replicação. O repositório do projeto está disponível no *github* no *link*: `https://github.com/francoisbarbosa-wq/TME_Ensemble_CG`.

4.6 Considerações Finais

Este capítulo detalhou, de forma integrada, o conjunto de decisões metodológicas adotadas para a classificação dos tecidos do TME de CG, desde a preparação dos dados até a formação dos comitês. Iniciou-se com a divisão da base por meio de *k-fold* estratificado por classe e por *clusters* intra-classe, preservando a heterogeneidade morfológica do *dataset*. Em seguida, conduziram-se treinamentos com arquiteturas originais, modificações arquiteturais (com uso de diferentes estratégias de extração de *embeddings*) e um arranjo hierárquico voltado a pares de classes conflituosas.

Após isso, as dez melhores por *F1-score* foram ajustadas quanto a quantidade de neurônios nas camadas totalmente conectadas, a fim de encontrar a melhor configuração para que pudéssemos ter nossos classificadores candidatos a membros do comitê. Por fim, construíram-se *ensembles* locais e globais, com múltiplas estratégias de apuração, constituiu um desenho metodológico que visa reduzir a variância e explorar complementaridades entre modelos. No capítulo seguinte, são apresentados e discutidos os resultados, avaliando em que medida essas escolhas metodológicas contribuíram para melhorar a discriminação entre os diferentes componentes teciduais do TME.

5 Resultados e Discussão

Neste capítulo, apresentamos os resultados dos experimentos conduzidos com as abordagens exibidas na Seção 4.2. Em seguida, são detalhados os resultados obtidos com o *ensemble* de dois níveis. Em todos os experimentos, as métricas foram obtidas por meio da média e do desvio padrão dos valores obtidos nas 8 execuções da validação cruzada. Cada experimento a seguir será identificado com a abreviação da arquitetura utilizada (Tabela 2), seguida de "_" e uma informação referente ao tipo de abordagem utilizada. Optou-se por utilizar a *efn* como referência para as metodologias subsequentes por ter sido o modelo utilizado no estudo proposto por Lou *et. al.* (LOU et al., 2025), estabelecendo-se assim como o principal parâmetro comparativo para este trabalho.

5.1 Resultados do Estudo Original

Lou *et al.* (LOU et al., 2025) apresentam, nos materiais complementares de seu artigo, o resultado das métricas e o algoritmo utilizado para o treinamento dos modelos propostos. Com base nessas informações, o primeiro experimento deste trabalho consistiu na replicação do estudo realizado pelos autores. A Tabela 5 mostra os resultados comparativos entre as duas abordagens, que foi o nosso ponto de partida para a sequência dos experimentos.

Tabela 5 – Resultados do estudo de Lou et al. (2025) e da replicação deste estudo.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
efn_Lou et al. (2025)	71,2 ± 1,8	96,0 ± 1,0	71,0 ± 1,8	73,7 ± 1,3	71,2 ± 1,8
efn_replicação	69,5 ± 2,2	95,2 ± 0,5	69,3 ± 1,9	72,3 ± 0,7	69,5 ± 2,2

-Todos os valores estão expressos em porcentagem.

Os resultados apresentados pelos autores do estudo original exibem valores superiores, ao experimento replicado, em todas as métricas. Quanto à estabilidade dos experimentos, ambos apresentam alto desvio padrão, sugerindo assim que em algumas das execuções, a replicação pode ter igualado ou superado os valores de Lou et al. (2025). Nas abordagens a seguir, buscamos melhorar tanto os valores das métricas quanto a estabilidade do modelo, reduzindo o desvio padrão entre as execuções.

5.2 Resultados com Arquitetura original

A Tabela 6 apresenta os resultados obtidos pelas arquiteturas originais, sem qualquer modificação estrutural, considerando apenas a preparação do *dataset* e os ajustes dos hiperparâmetros. O modelo *rsn101* apresentou os melhores resultados em todas as métricas, assim como uma alta estabilidade, refletida na baixa variação entre as execuções. Em seguida, *dsn121* e *efn_b2* mantiveram resultados próximos, também com baixa variância entre execuções.

Tabela 6 – Resultados comparativos dos modelos sem alterações na arquitetura.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
rsn101	**75,9 ± 0,4**	**96,8 ± 0,1**	**75,9 ± 0,4**	76,0 ± 0,4	**75,9 ± 0,4**
dsn121	75,2 ± 0,4	96,3 ± 0,2	75,2 ± 0,4	75,4 ± 0,4	75,2 ± 0,4
efn_b2	74,9 ± 0,3	96,2 ± 0,1	74,8 ± 0,3	75,1 ± 0,3	74,9 ± 0,3
efn	74,6 ± 0,4	96,2 ± 0,1	74,6 ± 0,4	74,7 ± 0,4	74,6 ± 0,4
rsn18	73,9 ± 0,3	96,0 ± 0,1	73,9 ± 0,3	74,0 ± 0,3	73,9 ± 0,3
dsn201	74,0 ± 0,8	96,2 ± 0,2	73,9 ± 0,7	74,7 ± 0,6	74,0 ± 0,8
mbn	73,8 ± 0,4	95,8 ± 0,1	73,8 ± 0,5	74,0 ± 0,4	73,8 ± 0,4
vit	73,3 ± 0,7	96,2 ± 0,1	73,2 ± 0,8	74,8 ± 0,8	73,3 ± 0,7
xcp	73,3 ± 0,6	95,1 ± 0,1	73,2 ± 0,5	73,3 ± 0,5	73,3 ± 0,6
vgg	72,6 ± 1,0	95,6 ± 0,2	72,5 ± 1,0	73,6 ± 0,7	72,6 ± 1,0
rsn50	71,3 ± 2,0	95,6 ± 0,6	71,2 ± 2,3	73,7 ± 1,0	71,3 ± 2,0
alx	70,0 ± 0,7	95,2 ± 0,1	69,8 ± 0,8	70,8 ± 0,8	70,0 ± 0,7

-Todos os valores estão expressos em porcentagem.

Modelos mais leves, em termos de parâmetros treináveis (conforme apresentado na Tabela 2), como *efn* e *mbn*, também apresentaram desempenhos satisfatórios, com F1-*Score* de 74,6% e 73,8%, respectivamente, embora o mbn tenha obtido AUC ligeiramente inferior a do experimento realizado pelos autores da base de dados. Redes como *rsn18* e o *dsn201* tiveram desempenhos ligeiramente similares porem dsn201 apresentou uma estabilidade inferior. O *vit* e o *xcp* , por sua vez, alcançaram *F1-score* e acurácia ligeiramente inferiores, sugerindo maior dificuldade em generalizar o conhecimento adquirido. Por outro lado, modelos maiores, como *vgg*, rsn50 e *alx* apresentaram piores resultados. Esses achados reforçam a relevância de arquiteturas mais modernas e profundas, como a *rsn101* e dsn121, no contexto de tarefas complexas de classificação multiclasse de componentes do TME.

Apesar dessa Fase incluir a avaliação de diferentes arquiteturas originais, as fases subsequentes (Fases 2, 3) foram conduzidas utilizando a *efn*, como mencionamos anteriormente. Essa escolha foi adotada para manter alinhamento metodológico com o trabalho de Lou *et al.* (LOU et al., 2025), que apresentou essa arquitetura como referência para o conjunto de dados, favorecendo uma comparação descritiva mais consistente do impacto

das modificações arquiteturais.

5.3 Resultados com Foco nos Vetores de Características

Nesta fase dos experimentos, utilizou-se a abordagem descrita na Seção 4.2.2, que consiste na extração de características em diferentes camadas convolucionais, usando o GAP para a construção de *embeddings* que serão concatenados posteriormente. Essa metodologia foi aplicada de três maneiras distintas, como podemos observar na Tabela 7: a primeira, *efn_GAP4* (Figura 21-b), utiliza quatro vetores concatenados; a segunda, *efn_GAP5* (Figura 21-c), utiliza cinco vetores; e, por fim, *efn_GAP2* (Figura 21-d), que usa apenas dois vetores, sendo este o que obteve o melhor desempenho entre os três.

Tabela 7 – Resultados de modificações arquiteturais usando GAPs concatenados.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
efn_GAP4	74,1 ± 0,3	96,2 ± 0,1	74,1 ± 0,4	74,4 ± 0,3	74,1 ± 0,3
efn_GAP5	74,2 ± 0,5	96,1 ± 0,2	74,2 ± 0,5	74,5 ± 0,4	74,2 ± 0,5
efn_GAP2	**74,8 ± 0,5**	**96,2 ± 0,1**	**74,7 ± 0,5**	**75,0 ± 0,5**	**74,8 ± 0,5**

-Todos os valores estão expressos em porcentagem.

A ilustração da Figura 37 mostra a aplicação do t-SNE (*t-distributed Stochastic Neighbor Embedding*) na projeção bidimensional dos *embeddings* extraídos em diferentes estágios das camadas convolucionais. A rede efn_GAP4 utiliza todos os vetores, exceto o referente à Figura 37-a. A efn_GAP5, por sua vez, utiliza os cinco vetores disponíveis, enquanto a efn_GAP2, que usa apenas os referentes às Figuras 37-d e 37-e.

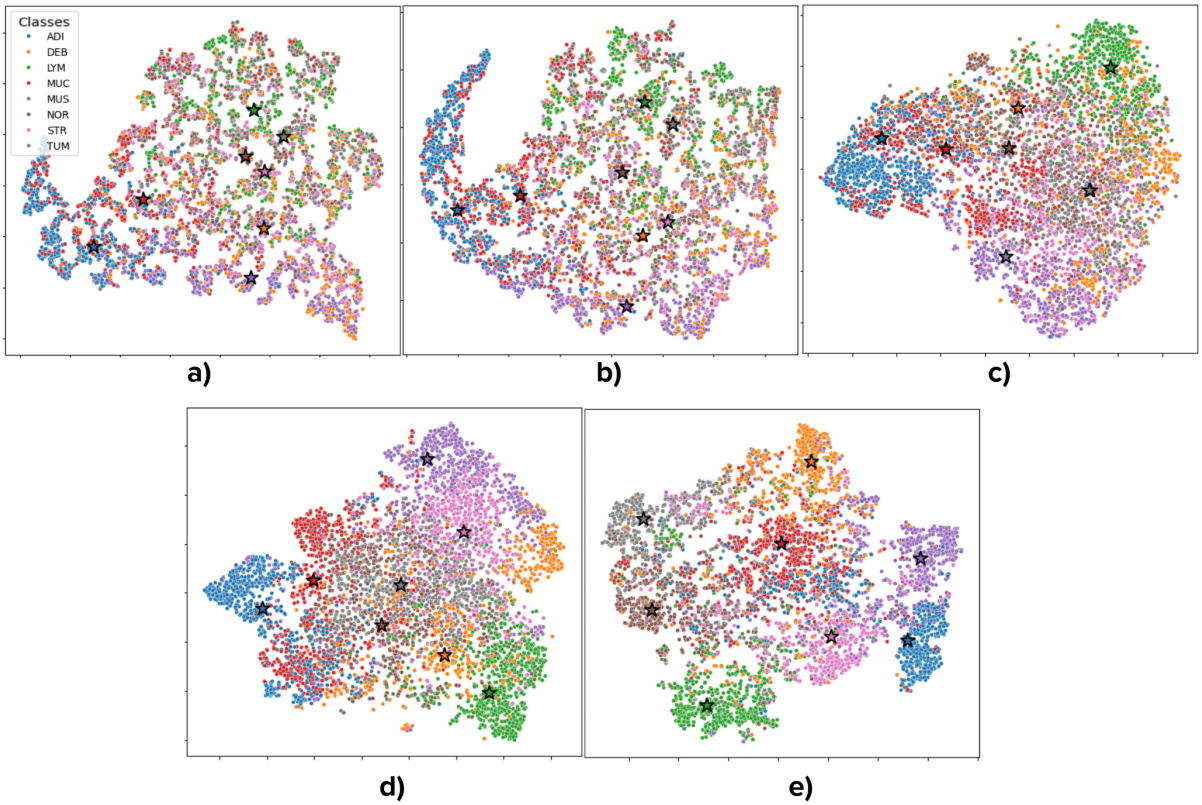


Figura 37 – Projeção bidimensional dos GAPs antes da concatenação. a) Vetores após a primeira convolução; b) apos a segunda convolução; c) após a quarta convolução; d) após a sexta convolução e e) após todas as convolução.

Aqui podemos observar com maior clareza os desafios referente a classificação multiclasse, sobretudo no contexto de imagens do TME do CG. Observa-se que, ao longo das convoluções, algumas classes apresentam uma sobreposição significativa em suas representações no espaço latente, estando mais próximas do centróide de outras classes. Destacam-se, por exemplo, as classes NOR e TUM, que apresentam agrupamentos com sobreposição considerável, evidenciando a dificuldade das redes em classificar amostras dessas classes.

5.4 Resultados com Hierarquização de treinamentos

Diante das dificuldades na separabilidade de classes morfologicamente semelhantes, a Tabela 8 apresenta os resultados de uma abordagem hierárquica de classificação proposta na Seção 5.4. Os modelos apresentados representam as duas variantes propostas dessa abordagem, onde a efn_h1 tem como foco realizar classificações binárias entre os pares mais conflitantes (ADIxMUC, MUSxSTR, e NORxTUM) sempre que um desses pares for predito pelo classificador multiclasse . Já a rede efn_h2 (Figura 22) hierarquizou toda a classificação de forma binária criando uma sequência de predições. Os resultados das todas métricas desses classificadores binários encontram-se no Anexo A.

Tabela 8 – Resultados comparativos do uso de classificadores hierárquicos .

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
efn_h1	**74,8 ± 0,3**	**95,1 ± 0,2**	**74,8 ± 0,3**	**74,9 ± 0,3**	**74,8 ± 0,3**
efn_h2	68,7 ± 0,5	82,1 ± 0,3	68,7 ± 0,5	69,7 ± 0,4	68,7 ± 0,5

-Todos os valores estão expressos em porcentagem.

O resultado de efn_h1 apresentou F1-*Score* satisfatório, aliado a uma boa estabilidade, evidenciado pelo baixo desvio padrão. No entanto, quanto à AUC, teve desempenho limitado, ficando com os mesmos valores do modelo de replicação (efn_replicação). Por outro lado, a rede efn_h2 obteve os piores resultados, sugerindo que a classificação hierárquica não foi eficaz para lidar com as ambiguidades entre classes.

5.5 Resultados Focados em Quantidade de Neurônios

Nas seções anteriores, foram apresentadas abordagens que utilizam a rede com sua arquitetura original inalterada, bem como abordagens com variações que modificam as camadas convolucionais. Para esta próxima fase dos experimentos, utilizaram-se alterações na quantidade de neurônios nas camadas totalmente conectadas e, para esta etapa, foram utilizados potenciais membros do comitê proposto, composto pelas oito melhores redes da Seção 5.2, a melhor da Seção 5.3 e a melhor da Seção 5.4, conforme a Tabela 9.

Tabela 9 – Resultados do candidatos a membros do comitê.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
rsn101	**75,9 ± 0,4**	**96,8 ± 0,1**	**75,9 ± 0,4**	**76,0 ± 0,4**	**75,9 ± 0,4**
dsn121	75,2 ± 0,4	96,3 ± 0,2	75,2 ± 0,4	75,4 ± 0,4	75,2 ± 0,4
efn_b2	74,9 ± 0,3	96,2 ± 0,1	74,8 ± 0,3	75,1 ± 0,3	74,9 ± 0,3
efn_h1	74,8 ± 0,3	95,1 ± 0,2	74,8 ± 0,3	74,9 ± 0,3	74,8 ± 0,3
efn_GAP2	74,8 ± 0,5	96,2 ± 0,1	74,7 ± 0,5	75,0 ± 0,5	74,8 ± 0,5
efn	74,6 ± 0,4	96,2 ± 0,1	74,6 ± 0,4	74,7 ± 0,4	74,6 ± 0,4
rsn18	73,9 ± 0,3	96,0 ± 0,1	73,9 ± 0,3	74,0 ± 0,3	73,9 ± 0,3
dsn201	74,0 ± 0,8	96,2 ± 0,2	73,9 ± 0,7	74,7 ± 0,6	74,0 ± 0,8
mbn	73,8 ± 0,4	95,8 ± 0,1	73,8 ± 0,5	74,0 ± 0,4	73,8 ± 0,4
vit	73,3 ± 0,7	96,2 ± 0,1	73,2 ± 0,8	74,8 ± 0,8	73,3 ± 0,7

-Todos os valores estão expressos em porcentagem.

A Tabela 20 mostra os resultados dos experimentos com variação de neurônios a partir das redes selecionadas. A nomenclatura adotada segue o mesmo padrão das fases anteriores, porém agora é adicionada a indicação do número de neurônios utilizados. A análise comparativa mostra que a alteração desse parâmetro tem um impacto sutil no desempenho, porém relevante no contexto do problema estudado.

Tabela 10 – Resultados comparativos dos modelos com variação na quantidade de neurônios na camada totalmente conectada.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
rsn101	75,9 ± 0,4	**96,8 ± 0,1**	75,9 ± 0,4	76,0 ± 0,4	75,9 ± 0,4
rsn101_512	**76,1 ± 0,3**	96,7 ± 0,1	**76,1 ± 0,2**	**76,4 ± 0,2**	**76,1 ± 0,3**
rsn101_256	75,6 ± 0,3	96,5 ± 0,2	75,7 ± 0,2	76,2 ± 0,4	75,6 ± 0,3
dsn121	75,2 ± 0,4	96,3 ± 0,2	75,2 ± 0,4	75,4 ± 0,4	75,2 ± 0,4
dsn121_512	**75,7 ± 0,5**	**96,6 ± 0,1**	**75,7 ± 0,5**	**75,9 ± 0,6**	**75,7 ± 0,5**
dsn121_256	75,6 ± 0,3	96,5 ± 0,1	75,6 ± 0,3	75,7 ± 0,4	75,6 ± 0,3
efn_b2	**74,9 ± 0,3**	96,2 ± 0,1	**74,8 ± 0,3**	75,1 ± 0,3	**74,9 ± 0,3**
efn_b2_512	74,7 ± 0,4	96,3 ± 0,1	74,7 ± 0,4	**75,2 ± 0,3**	74,7 ± 0,4
efn_b2_256	74,8 ± 0,4	**96,3 ± 0,1**	74,8 ± 0,4	75,1 ± 0,4	74,8 ± 0,4
efn_h1	**74,8 ± 0,3**	**95,1 ± 0,2**	**74,8 ± 0,3**	**74,9 ± 0,3**	**74,8 ± 0,3**
efn_h1_512	74,7 ± 0,4	95,0 ± 0,1	74,7 ± 0,4	74,9 ± 0,4	74,7 ± 0,4
efn_h1_256	74,6 ± 0,3	95,0 ± 0,1	74,5 ± 0,4	74,7 ± 0,4	74,6 ± 0,3
efn_GAP2	**74,8 ± 0,5**	**96,2 ± 0,1**	**74,7 ± 0,5**	**75,0 ± 0,5**	**74,8 ± 0,5**
efn_GAP2_512	74,1 ± 0,5	96,1 ± 0,2	74,1 ± 0,5	74,5 ± 0,4	71,1 ± 0,5
efn_GAP2_256	73,2 ± 0,5	95,9 ± 0,2	73,3 ± 0,4	73,7 ± 0,4	73,2 ± 0,5
efn	**74,6 ± 0,4**	**96,2 ± 0,1**	**74,6 ± 0,4**	**74,7 ± 0,4**	**74,6 ± 0,4**
efn_512	74,5 ± 0,4	96,2 ± 0,1	74,4 ± 0,4	74,7 ± 0,4	74,5 ± 0,4
efn_256	74,5 ± 0,5	96,1 ± 0,1	74,4 ± 0,5	74,5 ± 0,5	74,5 ± 0,5
rsn18	**73,9 ± 0,3**	**96,0 ± 0,1**	**73,9 ± 0,3**	**74,0 ± 0,3**	**73,9 ± 0,3**
rsn18_512	71,8 ± 1,0	95,5 ± 0,4	71,6 ± 1,1	72,8 ± 0,9	71,8 ± 1,0
rsn18_256	72,8 ± 0,8	95,7 ± 0,2	72,7 ± 0,8	73,5 ± 0,5	72,8 ± 0,8
dsn201	74,0 ± 0,8	96,2 ± 0,2	73,9 ± 0,7	74,7 ± 0,6	74,0 ± 0,8
dsn201_512	**74,1 ± 1,0**	**96,3 ± 0,2**	**74,0 ± 1,1**	**74,9 ± 1,0**	**74,1 ± 1,0**
dsn201_256	73,4 ± 1,0	96,2 ± 0,3	73,2 ± 1,0	74,4 ± 0,7	73,4 ± 1,0
mbn	73,8 ± 0,4	95,8 ± 0,1	73,8 ± 0,5	74,0 ± 0,4	73,8 ± 0,4
mbn_512	73,5 ± 0,8	95,8 ± 0,3	73,5 ± 0,8	74,2 ± 0,5	73,5 ± 0,8
mbn_256	**74,1 ± 0,4**	**96,1 ± 0,1**	**74,0 ± 0,3**	**74,4 ± 0,3**	**74,1 ± 0,4**
vit	73,3 ± 0,7	96,2 ± 0,1	73,2 ± 0,8	74,8 ± 0,8	73,3 ± 0,7
vit_512	**74,6 ± 0,6**	**96,2 ± 0,1**	**74,5 ± 0,6**	**75,0 ± 0,6**	**74,6 ± 0,6**
vit_256	73,1 ± 0,8	96,2 ± 0,2	73,0 ± 0,8	74,1 ± 0,8	73,1 ± 0,8

- Todos os valores estão expressos em porcentagem.

Para a arquitetura rsn101 e dsn121, as configurações com 512 neurônios (rsn101_512 e dsn121_512) obtiveram desempenho superior em quase todas as métricas, em comparação com as configurações base e as versões com 256 neurônios. Especificamente, a rsn101_512 atingiu *F1-score* de 76,1%, enquanto o AUC foi de 96,7%, sendo superado apenas pela rsn101, com 97,2%. Por outro lado, o resultado do modelo dsn121_512 apresenta os melhores resultados entre as variações de neurônios, porém ficando atrás apenas

da rsn101_512 com *F1-score* de 75,7%.

A arquitetura original da efn_b2 e a variação hierarquizada (efn_h1) apresentaram desempenho equivalente, com *F1-score* superior ao das demais configurações avaliadas com EfficientNet. A rede efn, tanto na sua configuração original quanto utilizando a concatenação de dois vetores de características (efn_GAP2), apresentou bons resultados, com uma pequena vantagem para a segunda. Em todos os casos, a versão base das redes teve o melhor desempenho em todas as métricas, superando os experimentos que envolviam alteração na quantidade de neurônios.

Dentre as redes escolhidas que apresentaram os piores desempenho temos, rsn18, dsn201_512 e mbn_256 quem apresentaram resultados semelhantes, destacando-se a mbn_256 pela maior estabilidade no *F1-score*. Esses resultados sugerem que diminuir a quantidade de neurônios para dsn201_512 e mbn_256 teve efeito positivo, mesmo diante da complexidade do problema.

Por fim, a arquitetura do ViT apresentou, em todas as métricas, seu melhor resultado na configuração de 512 neurônios (vit_512), alcançando 74,6% de acurácia e AUC de 96,2%. Esses resultados mostram que, mesmo em modelos baseados em mecanismos de atenção, modificar a quantidade de neurônios exerce papel relevante, mesmo em tarefas complexas de classificação multiclasse.

5.6 Resultados com *Ensemble* de dois Níveis

Nesta seção, apresentaremos os resultados obtidos com a metodologia proposta, porém, antes de montarmos o comitê, foi realizado o aumento de dados com as transformações tratadas na Seção 4.2.5 com objetivo de aumentar a variabilidade das amostras e melhorar a capacidade de generalização dos modelos. A Tabela 11 apresenta uma comparação direta entre os resultados dos modelos treinados com e sem o uso dessas técnicas de aumento.

O modelo efn foi o que mais se beneficiou do aumento de dados, apresentando ganhos expressivos de aproximadamente 2,5 pontos percentuais na acurácia e no *F1-score*. Por outro lado, o vit_512 foi o modelo que teve menor sensibilidade a essa estratégia, mantendo os resultados praticamente inalterados. Contudo, estes resultados mostram a importância do aumento de dados em tornar o modelo mais robusto a variações reais que possam ocorrer em novos dados.

Tabela 11 – Resultados dos modelos com e sem aumento de dados.

Aumento	Modelo	Acurácia	AUC	F1-*Score*	Precisão	*Recall*
Não	rsn101_512	76,1 ± 0,3	96,7 ± 0,1	76,1 ± 0,2	76,4 ± 0,2	76,1 ± 0,3
	dsn121_512	75,7 ± 0,5	96,6 ± 0,1	75,7 ± 0,5	75,9 ± 0,6	75,7 ± 0,5
	efn_b2	74,9 ± 0,3	96,2 ± 0,1	74,8 ± 0,3	75,1 ± 0,3	74,9 ± 0,3
	efn_h1	74,8 ± 0,3	95,1 ± 0,2	74,8 ± 0,3	74,9 ± 0,3	74,8 ± 0,3
	efn_GAP2	74,8 ± 0,5	96,2 ± 0,1	74,7 ± 0,5	75,0 ± 0,5	74,8 ± 0,5
	efn	74,6 ± 0,4	96,2 ± 0,1	74,6 ± 0,4	74,7 ± 0,4	74,6 ± 0,4
	vit_512	**74,6 ± 0,6**	96,1 ± 0,1	74,5 ± 0,6	75,0 ± 0,6	**74,6 ± 0,6**
	mbn_256	74,1 ± 0,4	96,1 ± 0,1	74,0 ± 0,3	74,4 ± 0,3	74,1 ± 0,4
	dsn201_512	74,1 ± 1,0	96,3 ± 0,2	74,0 ± 1,1	74,9 ± 1,0	74,1 ± 1,0
	rsn18	73,9 ± 0,3	96,0 ± 0,1	73,9 ± 0,3	74,0 ± 0,3	73,9 ± 0,3
Sim	rsn101_512	**77,5 ± 0,2**	**97,4 ± 0,1**	**77,5 ± 0,2**	**77,9 ± 0,2**	**77,5 ± 0,2**
	dsn121_512	**77,5 ± 0,4**	**97,3 ± 0,1**	**77,5 ± 0,4**	**77,8 ± 0,4**	**77,5 ± 0,4**
	efn_b2	**76,6 ± 0,6**	**97,0 ± 0,1**	**76,6 ± 0,6**	**76,8 ± 0,6**	**76,6 ± 0,6**
	efn_h1	**77,1 ± 0,3**	**95,9 ± 0,1**	**77,1 ± 0,3**	**77,2 ± 0,3**	**77,1 ± 0,3**
	efn_GAP2	**76,0 ± 0,6**	**97,0 ± 0,1**	**75,9 ± 0,6**	**76,2 ± 0,5**	**76,0 ± 0,5**
	efn	**77,1 ± 0,4**	**97,2 ± 0,1**	**77,2 ± 0,4**	**77,4 ± 0,3**	**77,1 ± 0,4**
	vit_512	74,5 ± 0,5	**96,6 ± 0,1**	**74,5 ± 0,5**	**75,0 ± 0,5**	74,5 ± 0,5
	mbn_256	**74,9 ± 0,7**	**96,8 ± 0,1**	**75,0 ± 0,7**	**75,7 ± 0,6**	**74,9 ± 0,7**
	dsn201_512	**76,9 ± 0,5**	**97,2 ± 0,1**	**77,0 ± 0,4**	**77,3 ± 0,4**	**76,9 ± 0,5**
	rsn18	**76,4 ± 0,5**	**97,0 ± 0,1**	**76,4 ± 0,5**	**76,8 ± 0,6**	**76,4 ± 0,5**

- Todos os valores estão expressos em porcentagem.

5.6.1 Nível 1: *Ensemble* Local

Nesta etapa, a Tabela 12 apresenta o desempenho dos comitês locais, que, como detalhado na Seção 4.4, consistem em combinar as probabilidades obtidas após a ativação *softmax* em cada execução da validação cruzada no split de teste de uma mesma arquitetura. Neste caso, cada arquitetura combinou utilizando duas estratégias de apuração: média das probabilidades e votação, ambas sem ponderações. De maneira geral, observa-se que, em comparação aos valores obtidos com aumento de dados, estas abordagens, resultaram em melhorias consistentes em todas as métricas avaliadas.

Ao comparar as estratégias de apuração, observou-se que a média das probabilidades teve uma ligeira superioridade em relação à votação, apresentando melhores resultados em todas as métricas avaliadas. Um ponto importante é que apenas a combinação por média possibilita o cálculo da AUC, uma vez que ela depende das probabilidades contínuas atribuídas a cada classe, enquanto a estratégia de votação considera apenas a classe predita, não fornecendo as probabilidades necessárias para a obtenção da AUC. Deste modo, optamos pelo uso das predições combinadas por média para a formação dos *ensembles* globais.

Tabela 12 – Resultados dos *ensembles* locais usando apuração por média e por votação, a partir das predições com aumento de dados

Combinação	**Modelo**	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
Média	rsn101_512	**80.6**	**98.0**	**80.6**	**80.8**	**80.6**
	dsn121_512	**80.0**	**97.9**	**80.1**	**80.2**	**80.1**
	dsn201_512	**79.4**	**97.8**	**79.5**	**79.7**	**79.4**
	efn	**79,1**	**97,6**	**79,1**	**79,3**	**79,1**
	efn_h1	**79.0**	**97.0**	**79.0**	**79.2**	**79.0**
	efn_b2	**79.0**	**97.6**	**79.0**	**79.1**	**79.0**
	rsn18	**78.8**	**97.6**	**78.8**	**79.1**	**78.8**
	efn_GAP2	**78,5**	**97,6**	**78,5**	**78,7**	**78,5**
	mbn_256	**78,4**	**97,5**	**78,5**	**78,8**	**78,4**
	vit_512	**78,0**	**97,3**	**78,0**	**78,1**	**78,0**
Votação	rsn101_512	80,2	–	80,2	80,2	80,2
	dsn121_512	79.8	–	79.8	79.9	79.8
	dsn201_512	79.3	–	79.3	79.5	79.3
	efn	79,0	–	79,0	79,0	79,0
	efn_h1	79.0	–	79.0	79.1	79.0
	efn_b2	78.7	–	78.7	78.9	78.7
	rsn18	78.6	–	78.5	78.7	78.6
	efn_GAP2	78,1	–	78,1	78,1	77,1
	mbn_256	77,9	–	78,0	78,3	77,9
	vit_512	77,3	–	77,3	77,3	77,3

- Todos os valores estão expressos em porcentagem.

5.6.2 Nivel 2: *Ensemble* Global

Após a definição dos modelos base mais promissores, aplicação do aumento de dados e criação dos comitês locais, foi implementado o *ensemble* global, que corresponde ao segundo nível da estratégia proposta. Esta etapa tem como objetivo usar as probabilidades combinadas dos comitês locais para montar comitês globais compostos por três, cinco ou sete redes, escolha que tentou evitar empate. As estrategias usadas para combinar suas probabilidades foram as descritas na Seção 4.4.3 e apresentada na Tabela 4.

Ao total foram montados 4.428 comitês diferentes, sendo 120 para cada estratégia com três e sete membros e 252 para cada estratégia com cinco membros, respeitando a formula de combinação simples (5.1), onde m é a quantidade de membros do comitê e r é o total de redes disponíveis. A Tabela 13 mostra as combinações com melhores resultados, para cada estratégia e cada quantidade de membros e suas respectivas abreviações.

$$\binom{r}{m} = C(r,m) = \frac{r!}{m!\,(r-m)!}, \qquad \text{para } 0 \leq m \leq r. \tag{5.1}$$

Tabela 13 – Combinações de classificadores e suas respectivas abreviações.

Método	Comitê	Combinação de Modelos
I	c1	(dsn121_512 , rsn101_512 ,efn_b2)
	c2	(dsn121_512 ,rsn101_512 , efn_GAP2 , vit_512 , efn_h1)
	c3	(dsn121_512 , rsn101_512 , rsn18 , efn_GAP2 , mbn_256 , vit_512 , efn_b2)
II	c4	(dsn121_512 , rsn101_512 , efn)
	c5	(dsn121_512 , rsn101_512 , efn , mbn_256 , vit_512)
	c6	(dsn121_512 ,rsn101_512 , rsn18 , efn_GAP2 , mbn_256 , vit_512 , efn_b2)
III	c7	(dsn121_512 , rsn101_512 , vit_512)
	c8	(dsn121_512 , rsn101_512 , efn , efn_h1 , dsn201_512)
	c9	(dsn121_512 , rsn101_512 , efn , efn_GAP2 , efn_h1 , efn_b2 , dsn201_512)
IV	c10	(dsn121_512 , rsn101_512 , efn_h1)
	c11	(dsn121_512 , rsn101_512 , mbn_256 , vit_512 , efn_h1)
	c12	(dsn121_512 , rsn101_512 , mbn_256 , vit_512 , efn_h1 , efn_b2 , dsn201_512)
V	c13	(dsn121_512 , rsn101_512 , efn)
	c14	(dsn121_512 , rsn101_512 , efn , rsn18 , mbn_256)
	c15	(dsn121_512 , rsn101_512 , efn , rsn18 , efn_GAP2 , mbn_256 , vit_512)
VI	c16	(dsn121_512 , rsn101_512 , vit_512)
	c17	(dsn121_512 , rsn101_512 , efn , rsn18 , efn_GAP2)
	c18	(rsn101_512 , efn , efn_GAP2 , mbn_256 , efn_h1 , efn_b2 , dsn201_512)
VII	c19	(dsn121_512 , rsn101_512 , efn_GAP2)
	c20	(dsn121_512 , rsn101_512 , mbn_256 , vit_512 , efn_b2)
	c21	(dsn121_512 , rsn101_512 , efn , rsn18 , mbn_256 , vit_512 , efn_h1)
VIII	c22	(dsn121_512 , rsn101_512 , efn)
	c23	(dsn121_512 ,rsn101_512 , efn , mbn_256 , vit_512)
	c24	(dsn121_512 , rsn101_512 , efn , rsn18 , mbn_256 , vit_512 , efn_h1)
IX	c25	(dsn121_512 , rsn101_512 , efn_b2)
	c26	(dsn121_512 , rsn101_512 , efn , rsn18 , efn_b2)
	c27	(dsn121_512 , rsn101_512 , efn , rsn18 , efn_GAP2 , efn_h1 , efn_b2)

A Tabela 14 apresenta os resultados dos comitês com melhores desempenhos em cada estratégia. Observa-se que a maioria das combinações resultou em melhorias em relação ao comitê local, com destaque para a $c20$ (dsn121_512 , rsn101_512 , mbn_256 , vit_512 , efn_b2), com 81,2% de acurácia, 98,1% de AUC e 81,2% de *F1-score*, evidenciando que o comitê global, usando a Estratégia VII que pondera a apuração com os *F1-score* de cada classe, foi capaz de compensar erros individuais, melhorando assim a predição final.

Tabela 14 – Resultados comparativos das combinações de classificadores.

Método	Combinação	Acurácia	AUC	F1-*Score*	Precisão	*Recall*
I	c1	80,9	98,1	80,9	81,1	80,9
	c2	81,1	98,0	81,1	81,2	81,1
	c3	81,0	98,1	81,0	81,1	81,0
II	c4	81,0	–	81,0	81,1	81,0
	c5	81,0	–	81,1	81,2	81,0
	c6	80,9	–	80,9	81,1	80,9
III	c7	80,7	92,4	80,7	80,8	80,7
	c8	80,3	92,8	80,3	80,5	80,3
	c9	80,1	93,4	80,2	80,3	80,1
IV	c10	80,8	97,6	80,7	80,7	80,8
	c11	80,9	97,6	80,8	80,8	80,9
	c12	80,8	97,6	80,7	80,7	80,8
V	c13	81,0	–	81,1	81,2	81,0
	c14	81,1	–	81,1	81,3	81,1
	c15	81,1	–	81,1	81,3	81,1
VI	c16	80,8	92,4	80,7	80,8	80,8
	c17	80,4	92,9	80,3	80,4	80,4
	c18	80,1	93,5	80,1	80,1	80,1
VII	c19	81,0	98,1	81,0	81,2	81,0
	c20	**81,2**	**98,1**	**81,2**	**81,3**	**81,2**
	c21	80,9	98,0	80,9	81,0	80,9
VIII	c22	80,9	–	80,9	81,0	80,9
	c23	81,1	–	81,1	81,2	81,1
	c24	81,0	–	81,0	81,2	81,0
IX	c25	80,9	92,2	80,9	81,1	80,9
	c26	80,5	93,2	80,5	80,7	80,5
	c27	80,2	93,4	80,2	80,4	80,2

Uma abordagem comparativa à nossa proposta é exibida na Tabela 15, onde são apresentados os resultados do *ensemble* tradicional, que combina os resultados de cada execução do *k-fold* entre os diversos membros, extraindo a média e o desvio padrão das métricas. Os resultados mostram um desempenho superior do *ensemble* proposto, mesmo ajustando as médias ao desvio padrão da abordagem tradicional. Essa superioridade evidencia que explorar a diversidade dos modelos durante a validação cruzada e, em seguida, refinar a decisão combinando os resultados de múltiplos classificadores é uma solução viável e vantajosa.

Tabela 15 – Resultados comparativos entre *Ensemble* tradicional e o *Ensemble* proposto.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
ensemble (Tradicional)	79,1 ± 0,3	97,7 ± 0,1	79,1 ± 0,3	79,3 ± 0,4	79,1 ± 0,3
ensemble (Proposto)	**81,2**	**98,1**	**81,2**	**81,3**	**81,2**

-Todos os valores estão expressos em porcentagem.

Por fim, a Tabela 16 mostra a comparação do classificador apresentado por Lou *et al.* (LOU et al., 2025) e o modelo proposto neste trabalho. Nota-se que o *ensemble* de dois níveis supera significativamente a abordagem original em todas as métricas avaliadas, obtendo uma melhora em mais de dois pontos percentuais na AUC (98,1% contra 96,0%). Além disso, observa-se uma diferença expressiva na no F1-score (81,2% contra 71,0%), o que reforça que estratégia proposta é mais equilibrada.

Tabela 16 – Resultados comparativos entre os autores da base e o modelo proposto.

Modelo	**Acurácia**	**AUC**	**F1-*Score***	**Precisão**	***Recall***
efn_Lou et al. (2025)	71,2 ± 1,8	96,0 ± 1,0	71,0 ± 1,8	73,7 ± 1,3	71,2 ± 1,8
ensemble (Proposto)	**81,2**	**98,1**	**81,2**	**81,3**	**81,2**

-Todos os valores estão expressos em porcentagem.

A comparação classe a classe entre o modelo replicado (baseado nos autores originais da base de dados) e o modelo proposto revela melhorias consistente em todas as classes, como mostrado na Tabela 17. Nota-se que tanto as classes com desempenho inicialmente limitado quanto aquelas que já apresentavam bons resultados obtiveram ganhos adicionais com modelo proposto neste trabalho.

Tabela 17 – Resultados comparativos por classe entre os autores da base e o modelo proposto (ganho relativo em pontos percentuais).

Classe	**Modelo Replicado**			**Modelo proposto (Δ)**		
	Precisão	**Recall**	**F1-Score**	**Precisão**	**Recall**	**F1-Score**
ADI	80,9	85,9	83,2	+7,5	+2,0	+4,9
DEB	59,7	77,4	67,0	+20,5	+0,0	+11,7
LYM	81,8	81,5	80,7	+7,0	+9,9	+9,3
MUC	73,1	65,8	68,1	+6,0	+18,1	+13,3
MUS	79,7	70,9	73,6	+8,5	+7,0	+9,1
NOR	72,8	52,1	60,2	+0,6	+25,5	+15,3
STR	56,8	71,0	61,0	+20,3	+3,6	+14,9
TUM	73,4	51,7	60,2	+2,3	+27,1	+17,0

Valores do Δ representam o ganho relativo do modelo proposto em pontos percentuais (p.p.).

A Figura 38 evidencia as melhorias significativas de algumas classes obtidas pelo modelo proposto. Destaca-se a classe DEB, que apresentou um aumento de mais de 20 pontos percentuais na precisão e mais de 10 pontos no *F1-score*, mesmo mantendo *recall* inalterado. Esse padrão sugere uma redução substancial dos falsos positivos, com manutenção da da capacidade de identificação dos casos positivos. Por sua vez, classes como LYM, MUS e MUC tiveram ganhos notáveis em todas as métricas, obtendo aumento superior a 8 pontos percentuais, indicando que o modelo apresentou um comportamento mais equilibrado.

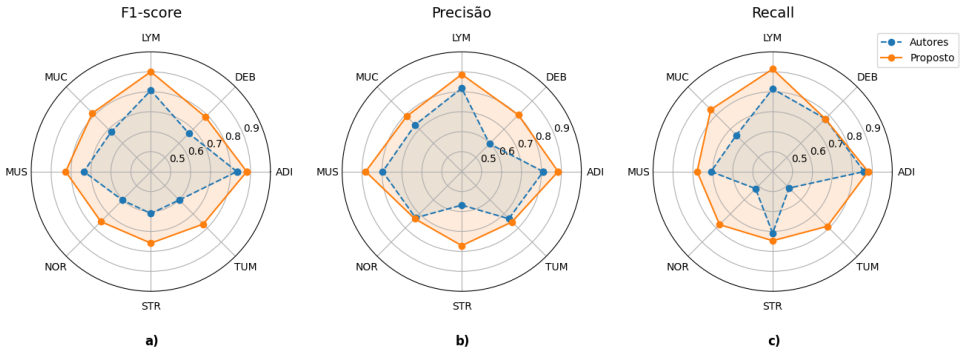


Figura 38 – Gráficos de radar comparando as métricas F1-score (a), Precisão (b) e Recall (c) entre o modelo proposto e os resultados dos autores da base.

Quando são analisadas classes mais problemáticas, como NOR e TUM, que inicialmente apresentavam *F1-score* de 60,2%, indicando alguma confusão com outras classes, observa-se que o modelo proposto apresentou uma melhora de mais de 17 pontos percentuais em ambas. Essa evolução demonstra uma maior capacidade discriminativa do novo modelo, especialmente entre tecidos morfologicamente semelhantes, como estruturas normais e tumorais. Outra classe crítica com melhorias significativas foi STR, cujo *F1-score* cresceu de 61,0% para 75,9%, reforçando a efetividade das melhorias aplicadas.

A matriz de confusão ilustrada na Figura 39 mostra que o modelo proposto melhorou a capacidade de separabilidade entre as principais classes conflituosas. Destacam-se as classes ADI e MUC, que no modelo replicado apresentavam alta ambiguidade entre si,com redução de 12% para 6% na taxa de confusão entre elas, além de melhorias em relação a outras classes com as quais anteriormente se confundiam.

As classes NOR e TUM são as que apresentavam as piores taxas de acerto, ambas inferiores 55% , apresentando elevados índices de conflito, seja entre elas ou com outras classes. A matriz revela melhora neste conflito, onde NOR, que era frequentemente predita como TUM (de 9% para 6%) e STR (de 13% para 4%), enquanto TUM era confundida principalmente com STR (de 17% para 5%) e DEB (de 12% para 3%). Com o modelo proposto, observa-se uma melhora substancial na taxa de acerto de ambas as classes, com NOR atingindo 78% de acerto e TUM subindo para 79%.

Esses resultados evidenciam as observações de Mikhailov *et al.* (MIKHAILOV; KHVOSTIKOV; KRYLOV, 2022), que o tecido estromal (STR) se encontra de forma difusa entre as células tumorais (TUM), dificultando assim sua separação, assim como a existência de semelhanças morfológicas entre tecidos normais e tumorais.

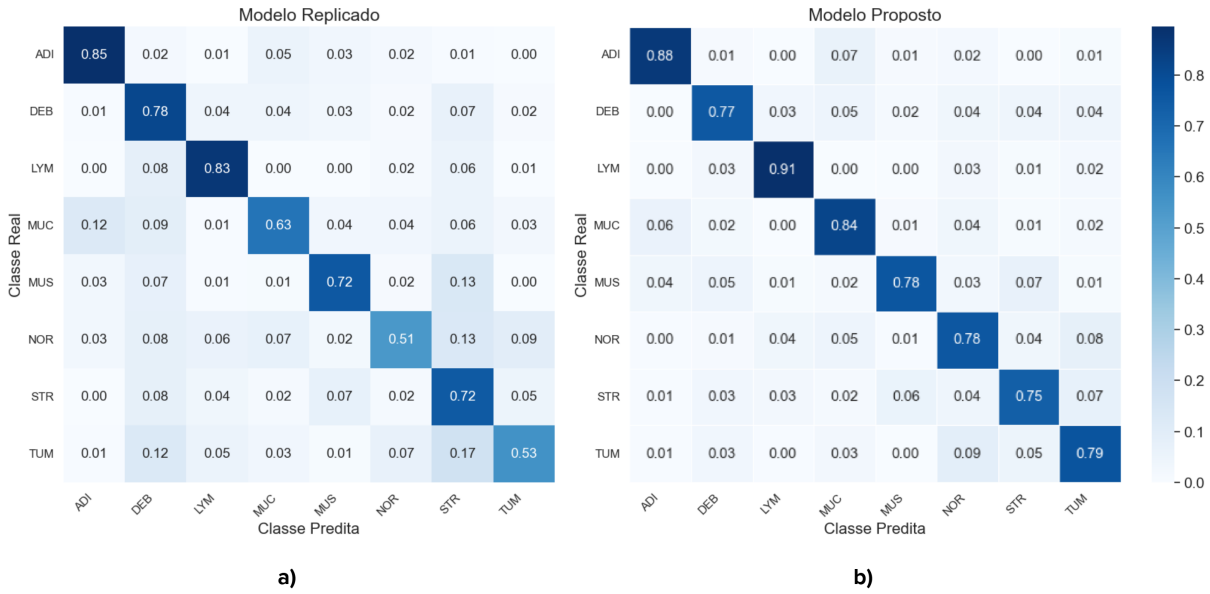


Figura 39 – Comparação entre as matrizes de confusão dos modelos replicado e proposto, destacando melhorias na separabilidade entre classes comumente confundidas.

A possibilidade de *overfitting* das redes que compõem o *ensemble* foi cuidadosamente investigada. Para isso, a Figura 40 apresenta gráficos da função de perda e da acurácia nos conjuntos de treino e validação ao longo das épocas. Em todos as avaliações posteriores (testes), os modelos utilizados foram aqueles salvos no menor ponto de *loss* de validação, diminuindo assim as chances de *overfitting*.

Após esse ponto, a condição de parada permitiu que o treinamento continuasse por até 10 épocas, desde que não houvesse melhora na função de perda. Os gráficos mostram que, até o ponto de menor *loss*, os conjuntos de treino e validação apresentam a mesma tendência de crescimento, garantindo que não houve indícios de *overfitting*.

Apesar das variações no comportamento da função de perda entre os diferentes modelos, todas as redes demonstraram capacidade de aprendizado até o ponto de menor *loss* de validação. As diferenças observadas na evolução das curvas refletem características próprias de cada arquitetura, como velocidade de convergência e sensibilidade à complexidade dos dados, não comprometendo a inclusão no *ensemble*. Ao contrário, essa diversidade de comportamento entre os modelos contribui positivamente para o comitê, promovendo complementaridade e aumento do poder de generalização, que é justamente um dos principais objetivos de estratégias baseadas em *ensemble*

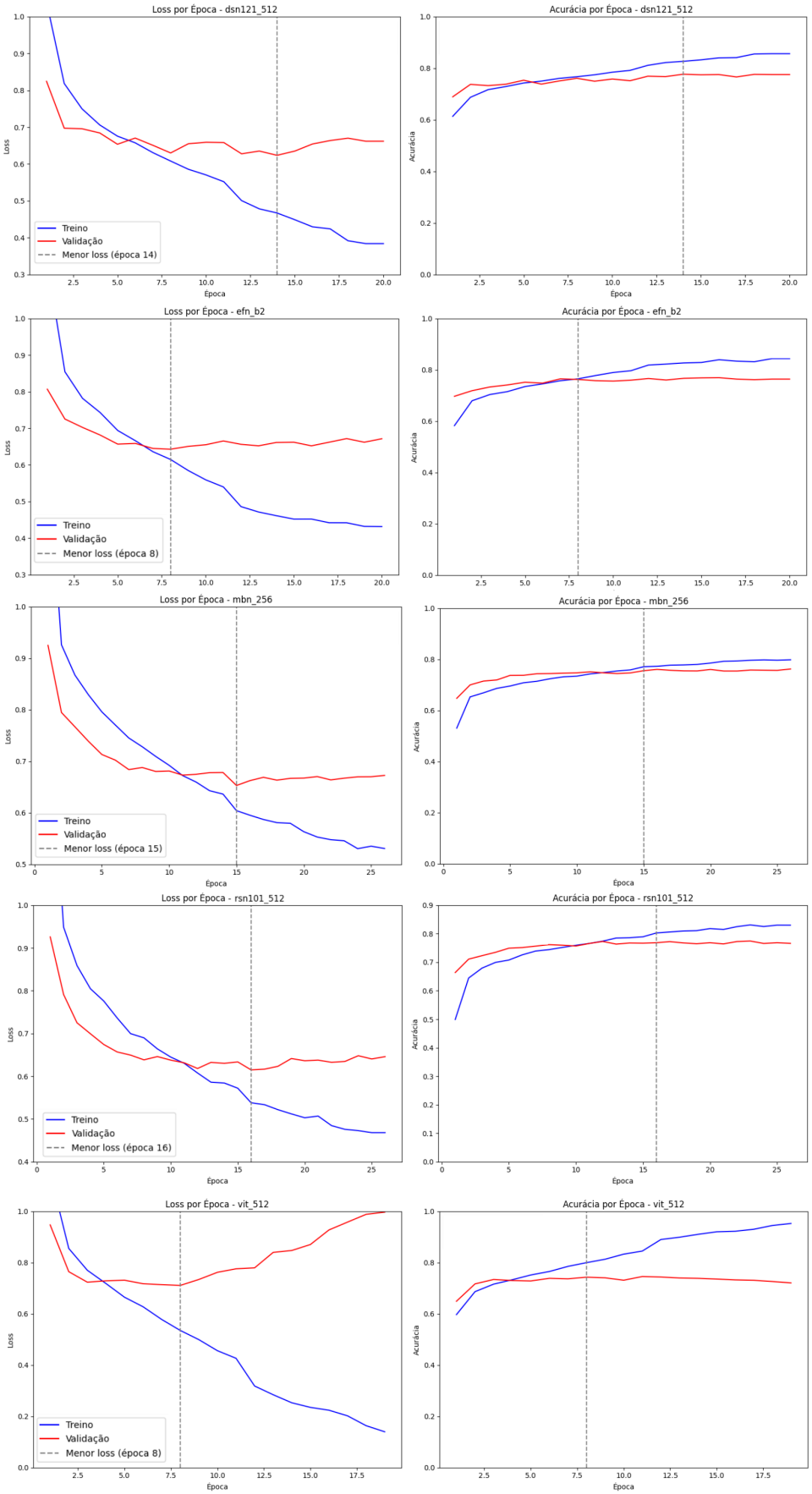


Figura 40 – *Loss* e acurácia das redes do ensemble c20 em cada época nos conjuntos de validação e treino

5.7 Explicabilidade

Com o finalidade de explicar as decisões do modelo, utilizou-se a técnica *Gradient-weighted Class Activation Map* (Grad-CAM) (SELVARAJU et al., 2016), que enfatiza em uma imagem as regiões mais relevantes para a predição de uma classe específica. O Grad-CAM gera mapas de ativação, destacando áreas que influenciam a decisão do modelo e oferece uma interpretação do que foi aprendido durante o treinamento.

Neste estudo, os mapas de ativação são sobrepostos às amostras, permitindo visualizar as regiões mais relevantes para a classificação dos tecidos do TME. A Figura 41 ilustra amostras classificadas pelo *ensemble* $c20$ de maneira correta, apresentando sua imagem original e os mapas gerados por cada modelo do comitê. Na borda superior das imagem, uma barra de cor indica se a classificação por aquela rede foi correta (verde) ou errônea (vermelha).

Observa-se que, nas predições incorretas, as regiões salientadas pelos modelos divergem substancialmente daquelas associadas às predições corretas. Por outro lado, quando analisamos somente as predições corretas por classe, vemos que os modelos focam em áreas complementares e ligeiramente coincidentes, mostrando que os padrões aprendidos apresentam convergência espacial relevante entre os classificadores.

Cabe ressaltar que essa interpretação é indicativa e demanda validação especializada. A análise por Grad-CAM deve ser interpretada com cautela, pois pode não refletir integralmente os critérios diagnósticos e precisariam ser validados por um patologista especializado. Contudo, o Grad-CAM não deve ser visto como explicação final do comportamento do modelo, uma vez que seus resultados podem não coincidir com o padrão clínico esperado.

5.8 Considerações Finais

Neste capítulo, foram apresentados e analisados os resultados obtidos nos diversos experimentos conduzidos ao longo do estudo, abrangendo tanto as arquiteturas em sua forma original quanto variações nas camadas convolucionais e nas totalmente conectadas, totalizando 37 abordagens. Em seguida, foram geradas 4.428 combinações distintas de comitês, sendo $c20$ o mais performático, superando de forma consistente tanto o classificador original de Lou *et al.* quanto o *ensemble* tradicional, elevando o *F1-score* global para 81,2% e a AUC para 98,1%, com ganhos expressivos em classes historicamente problemáticas. Esses resultados indicam que a combinação estruturada de arquiteturas diversas, aliada ao uso de comitês locais e globais, não apenas reduz a variância entre execuções, mas também melhora a capacidade discriminativa em contextos multiclasse complexos, explorando complementaridades entre modelos.

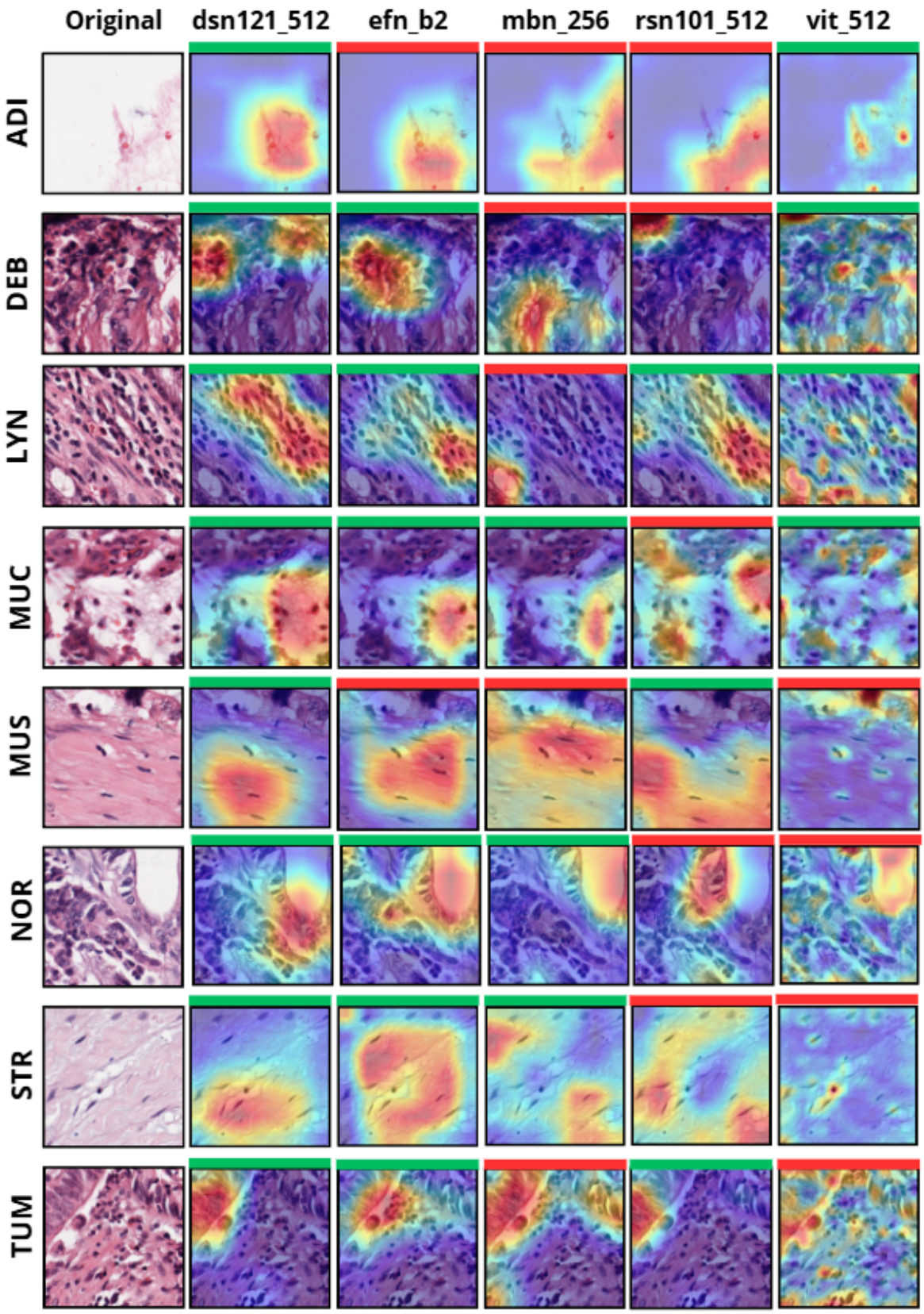


Figura 41 – Mapas ativação de exemplos das oito classes de tecido para os cinco modelos avaliados. A primeira coluna mostra a imagem original e a barra superior indica acertos (verde) e erros (vermelho).

6 Conclusões e Trabalhos Futuros

Este capítulo apresenta as considerações finais deste trabalho, destacando suas contribuições para o campo da patologia digital na tarefa de classificação de componentes do microambiente tumoral do câncer gástrico utilizando algoritmos de *Deep Learning*. Ao longo da pesquisa, com base na revisão bibliográfica, destacou-se a relevância do estudo nessa área, tanto para definir condutas terapêuticas mais precisas, quanto para desenvolver novos biomarcadores. Contudo, a tarefa de classificação é fundamental para que se utilizem os componentes adequados para que cada uma dessas abordagens seja eficaz.

Neste trabalho, foi desenvolvido um método de classificação de imagens dos diferentes tipos de tecidos do TME de CG. A proposta baseia-se em um comitê hierárquico contendo dois níveis, sendo um local, com *ensembles* homogêneos, onde os membros são os modelos oriundos do *k-fold*, e outro global, com os membros sendo os comitês locais. A seleção dos potenciais membros do comitê foi orientada por abordagens metodológicas consolidadas na literatura, incluindo concatenação de vetores de características, variação na quantidade de neurônios, aumento de dados e transferência de aprendizado. Ao todo, foram utilizadas trinta e sete arquiteturas diferentes de DL (levando em conta as originais e as modificadas).

Após a avaliação do desempenho das redes, foram selecionadas para a etapa de alteração da quantidade de neurônios as oito melhores redes dentre as arquiteturas sem alterações (originais) e a melhor em cada alteração de arquitetura. Em seguida as melhores modificações da camada totalmente conectada foram submetidas ao seu comitê local, nos quais as diferentes execuções da validação cruzada atuaram como membros de cada *ensemble*, combinados por média e por votação. A análise comparativa indicou que a combinação por média apresentou desempenho superior, sendo, portanto, adotada também na etapa de *ensemble* global. Nessa fase, foram geradas todas as combinações possíveis com três, cinco e sete membros nas diferentes estratégias de combinações, totalizando 4428 combinações distintas. O comitê com melhor desempenho foi formado pelas arquiteturas dsn121_512, efn_b2, mbn_256, rsn101_512, vit_512, resultando em melhorias expressivas em todas as métricas avaliadas em relação ao estudo original.

De forma sintética, este trabalho teve como objetivo desenvolver um comitê de redes neurais, em dois níveis, para a classificação multiclasse de imagens histológicas de tecidos do TME de CG, combinando predições para identificar as diferentes estruturas morfológicas que o compõem. Os resultados obtidos mostram que o comitê hierárquico proposto alcançou desempenho superior aos modelos individuais e ao estudo original de

referência, com melhorias consistentes em todas as métricas avaliadas, indicando que o objetivo geral foi atingido.

Desta forma, conclui-se que o uso de dois níveis de *ensemble* favorece a generalização do modelo, permitindo combinar diferentes perspectivas de treinamento, seja no nível local, na validação cruzada, seja no nível global, entre os modelos distintos. Essa estratégia possibilita a exploração complementar de diferentes pontos fortes das arquiteturas envolvidas, contribuindo assim para o desempenho global do modelo proposto.

Apesar dos resultados promissores, este estudo apresenta a limitação de ter usado apenas uma base de dados para CG, o que restringe a avaliação da capacidade de generalização do modelo proposto. Ainda sim, empregamos validação cruzada com *split* de teste externo e agrupamentos de *patches* para preservar a heterogeneidade do conjunto de treinamento, mitigando parcialmente esse risco.

Como extensões naturais deste trabalho, propõe-se: (i) avaliar o comitê hierárquico em novos *datasets* de CG provenientes de outras fontes, a fim de verificar sua capacidade de generalização frente à variabilidade entre centros hospitalares; (ii) explorar *ensembles* com método de *stacking*, são agregadas para formar um metaconjunto de treinamento, permitindo aprender pesos e interações de forma mais flexível do que a combinação por média ou votação; e por fim, (iii) aplicar o desenho metodológico destre trabalho em bases de outras patologias, a fim de avaliar se os ganhos observados neste estudo se estendem a diferentes tarefas de classificação em patologia digital.

Referências

ABDAR, M.; MAKARENKOV, V. Cwv-bann-svm ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, v. 146, p. 557–570, 2019. ISSN 0263-2241. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0263224119304385>. Citado na página 52.

BALLARD, D. H.; BROWN, C. M. *Computer Vision*. New York: Prentice Hall Professional Technical Reference, 1982. Citado na página 22.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 1251–1258. Citado na página 44.

de Visser, K. E.; JOYCE, J. A. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*, v. 41, n. 3, p. 374–403, 2023. ISSN 1535-6108. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1535610823000442>. Citado 2 vezes nas páginas 16 e 20.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. Disponível em: <http://dx.doi.org/10.1109/CVPR.2009.5206848>. Citado na página 44.

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 30.

DOSOVITSKIY, A. et al. An image is worth 16Œ16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*. Austria: [s.n.], 2021. Disponível em: <https://arxiv.org/abs/2010.11929>. Disponível em: <https://arxiv.org/abs/2010.11929>. Acesso em: 07 jul. 2025. Citado 3 vezes nas páginas 7, 28 e 29.

DUTSCHMANN, T.-M. et al. Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. *Journal of Cheminformatics*, v. 15, p. 49, 2023. Citado na página 51.

FERLAY, J. et al. *Global Cancer Observatory: Cancer Today*. Lyon, France: [s.n.], 2024. Site institucional da International Agency for Research on Cancer. Disponível em: <https://gco.iarc.who.int/today>. Acesso em: 19 mai. 2025. Citado 3 vezes nas páginas 7, 15 e 16.

GANAIE, M. et al. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, v. 115, p. 105151, 2022. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095219762200269X>. Citado na página 53.

GOCERI, E. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial intelligence review*, Springer, v. 56, n. 11, p. 12561–12605, 2023. Citado na página 29.

GUEHRIA, S.; BELLEILI, H.; AZIZI, N. A survey on ensemble multi-label classifiers. In: ABRAHAM, A. et al. (Ed.). *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)*. Cham: Springer Nature Switzerland, 2023, (Lecture Notes in Networks and Systems, v. 648). p. 100–109. Disponível em: <https://doi.org/10.1007/978-3-031-27524-1_11>. Acesso em: 20 mai. 2025. Citado na página 30.

GUIDO, S.; MüLLER, A. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Gravenstein Highway North: O'Reilly Media, 2017. Citado na página 23.

HAMEED, Z. et al. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. *Scientific Reports*, Nature Publishing Group, v. 12, n. 1, p. 15600, 2022. Disponível em: <https://doi.org/10.1038/s41598-022-19278-2>. Citado 2 vezes nas páginas 30 e 45.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2. ed. Porto Alegre: Bookman Editora, 2001. Citado 2 vezes nas páginas 23 e 24.

HE, K. et al. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, p. 770–778, 2015. Citado na página 25.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 44.

HOLZINGER, A. et al. *Artificial Intelligence and Machine Learning for Digital Pathology*. Cham: Springer International Publishing, 2020. v. 12090. ISBN 978-3-030-50401-4. Disponível em: <https://doi.org/10.1007/978-3-030-50402-1>. Citado na página 21.

HOWARD, A. et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019. Citado na página 44.

HOWARD, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. Disponível em: <https://arxiv.org/abs/1704.04861>. Citado na página 27.

HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 4700–4708. Citado 2 vezes nas páginas 26 e 44.

IGLOVIKOV, V. et al. *Albumentations Documentation*. 2025. <https://albumentations.ai/docs/>. Acesso em: 12 ago. 2025. Citado na página 48.

Instituto Nacional de Câncer (INCA). *Estômago Taxas ajustadas por 100 mil habitantes*. 2024. Acesso em: 9 jun. 2025. Disponível em: <https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/por-neoplasia-taxas-ajustadas/estomago>. Citado na página 15.

KATHER, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, v. 16, n. 1, p. e1002730, 2019. Disponível em: <https://doi.org/10.1371/journal.pmed.1002730>. Citado 6 vezes nas páginas 20, 35, 36, 37, 38 e 41.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 44.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. Citado 2 vezes nas páginas 7 e 24.

LIANG, P. et al. Enhancing whole slide image classification with discriminative and contrastive learning. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. [S.l.]: Springer Nature Switzerland, 2024. (Lecture Notes in Computer Science, v. 15004), p. 102–112. Citado na página 43.

LOU, S. et al. *Gastric Cancer Histopathology Tissue Image Dataset (GCHTID)*. figshare, 2024. Dataset. Disponível em: <https://doi.org/10.6084/m9.figshare.25954813.v1>. Citado 4 vezes nas páginas 7, 21, 41 e 45.

LOU, S. et al. A large histological images dataset of gastric cancer with tumour microenvironment annotation for ai. *Scientific Data*, v. 12, p. 138, 2025. Disponível em: <https://doi.org/10.1038/s41597-025-04489-9>. Acesso em: 19 mai. 2025. Citado 12 vezes nas páginas 7, 10, 16, 36, 37, 38, 39, 41, 42, 59, 60 e 70.

LUGER, G. F. *Inteligência Artificial*. São Paulo: Pearson Education do Brasil, 2013. Tradução: Daniel Vieira. Citado na página 21.

MANDAL, S. et al. The tumour histopathology glossary for ai developers. *PLOS Computational Biology*, Public Library of Science, v. 21, n. 1, p. e1012708, 2025. Disponível em: <https://doi.org/10.1371/journal.pcbi.1012708>. Citado 2 vezes nas páginas 17 e 20.

MIKHAILOV, I.; KHVOSTIKOV, A.; KRYLOV, A. Methodical approaches to annotation and labeling of histological images in order to automatically detect the layers of the stomach wall and the depth of invasion of gastric cancer. *Archive of Pathology = Arkhiv patologii*, v. 84, n. 6, p. 67–73, 2022. In Russian. Disponível em: <https://doi.org/10.17116/patol20228406167>. Citado 6 vezes nas páginas 21, 34, 36, 37, 45 e 72.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. *Array*, v. 16, p. 100258, 2022. ISSN 2590-0056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2590005622000911>. Citado na página 29.

RONG, R. et al. A deep learning approach for histology-based nuclei segmentation and tumor microenvironment characterization. *bioRxiv*, dec 2022. Preprint. Disponível em: <https://doi.org/10.1101/2022.12.08.519641>. Citado 3 vezes nas páginas 35, 37 e 38.

RUGGE, M. et al. Gastritis: The clinico-pathological spectrum. *Digestive and Liver Disease*, v. 53, n. 10, p. 1237–1246, 2021. Review Article. Citado na página 19.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Pearson, 2016. Always learning. ISBN 9781292153964. Citado na página 21.

SANTOS, J. et al. Glomerulosclerosis identification using a modified dense convolutional network. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2022. p. 237–252. Citado na página 48.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1610.02391>. Citado na página 74.

SERAG, A. et al. Translational ai and deep learning in diagnostic pathology. *Frontiers in Medicine*, v. 6, p. 185, 2019. Citado na página 17.

SHAIKH, H.; WEHRLE, C. J.; KHORASANI-ZADEH, A. *Anatomy, Abdomen and Pelvis: Mesenteric Artery*. 2023. <https://www.ncbi.nlm.nih.gov/books/NBK538323/>. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Citado na página 19.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, Springer, v. 6, p. 60, jul. 2019. Disponível em: <https://doi.org/10.1186/s40537-019-0197-0>. Citado na página 30.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado 3 vezes nas páginas 24, 27 e 44.

TAJBAKHSH, N. et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, Institute of Electrical and Electronics Engineers (IEEE), v. 35, n. 5, p. 12991312, maio 2016. ISSN 1558-254X. Disponível em: <http://dx.doi.org/10.1109/TMI.2016.2535302>. Citado na página 24.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. [s.n.], 2019. Disponível em: <https://arxiv.org/abs/1905.11946>. Disponível em: <https://arxiv.org/abs/1905.11946>. Acesso em: 22 jun. 2025. Citado 2 vezes nas páginas 28 e 44.

TURING, A. M. *Computing Machinery and Intelligence*. 1950. Disponível em: <https://www.csee.umbc.edu/courses/471/papers/turing.pdf>. Acesso em: 20 jun. 2025. Citado na página 22.

VASWANI, A. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2017. p. 6000–6010. Citado na página 28.

WANG, S. et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Research*, American Association for Cancer Research, v. 80, n. 10, p. 2056–2066, 2020. Published online ahead of print March 1, 2021. Citado 2 vezes nas páginas 35 e 37.

WANG, Z. et al. Identification of histopathological classification and establishment of prognostic indicators of gastric adenocarcinoma based on deep learning algorithm. *Medical Molecular Morphology*, v. 57, p. 286–298, 2024. Disponível em: <https://doi.org/10.1007/s00795-024-00399-8>. Acesso em: 19 mai. 2025. Citado 5 vezes nas páginas 17, 20, 36, 37 e 38.

ZHANG, X.; LIU, C.-A. Model averaging prediction by k-fold cross-validation. *Journal of Econometrics*, v. 235, n. 1, p. 280–301, 2023. Citado na página 51.

ZHENG, Y. et al. Hierarchical convolutional neural network via hierarchical cluster validity based visual tree learning. *Neurocomputing*, v. 409, p. 408–419, 2020. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092523122030970X>. Citado na página 46.

ZHOU, X. et al. Construction and validation of a deep learning prognostic model based on digital pathology images of stage iii colorectal cancer. *European Journal of Surgical Oncology*, Elsevier, v. 50, p. 108369, 2024. Disponível em: <https://doi.org/10.1016/j.ejso.2024.108369>. Citado 3 vezes nas páginas 36, 37 e 38.

ZHOU, Z.-H. *Ensemble Methods: Foundations and Algorithms*. [S.l.]: Taylor & Francis, 2012. (Chapman & Hall/CRC Machine Learning & Pattern Recognition Series). ISBN 9781439830031. Citado 2 vezes nas páginas 7 e 30.

Anexos

ANEXO A – Resultados dos Classificadores Binários

Tabela 18 – Resultados com variação na quantidade de neurônios dos classificadores binários, sem aumento de dados.

Modelo	Par	Acurácia	AUC	F1-*Score*	Precisão	*Recall*
efn		**88,9 ± 0,7**	**95,8 ± 0,2**	**88,9 ± 0,8**	**89,0 ± 0,7**	**88,9 ± 0,7**
efn_512	ADI x MUC	88,6 ± 0,6	95,8 ± 0,3	88,6 ± 0,6	88,8 ± 0,5	88,6 ± 0,6
efn_256		88,9 ± 0,7	95,7 ± 0,4	88,9 ± 0,6	89,1 ± 0,5	88,9 ± 0,6
efn		**87,6 ± 0,7**	95,1 ± 0,3	**87,6 ± 0,7**	**87,8 ± 0,6**	**87,6 ± 0,7**
efn_512	MUS x STR	87,3 ± 0,4	**95,1 ± 0,2**	87,3 ± 0,4	87,5 ± 0,4	87,3 ± 0,4
efn_256		87,5 ± 0,6	94,9 ± 0,2	87,5 ± 0,6	87,7 ± 0,5	87,5 ± 0,6
efn		84,4 ± 0,7	92,2 ± 0,2	84,4 ± 0,7	84,5 ± 0,7	84,4 ± 0,7
efn_512	NOR x TUM	**84,4 ± 0,5**	**92,1 ± 0,2**	**84,4 ± 0,5**	**84,5 ± 0,5**	**84,4 ± 0,5**
efn_256		84,1 ± 0,4	91,9 ± 0,2	84,0 ± 0,5	84,4 ± 0,4	84,1 ± 0,4

- Todos os valores estão expressos em porcentagem.

Tabela 19 – Resultados com dos melhores classificadores binários com aumento de dados.

Modelo	Par	Acurácia	AUC	F1-*Score*	Precisão	*Recall*
efn	ADI x MUC	90,1 ± 0,3	96,8 ± 0,1	90,1 ± 0,3	90,1 ± 0,3	90,1 ± 0,3
efn	MUS x STR	89,4 ± 0,2	96,5 ± 0,1	89,4 ± 0,2	89,5 ± 0,2	89,4 ± 0,2
efn_512	NOR x TUM	86,3 ± 0,6	94,1 ± 0,2	86,3 ± 0,6	86,4 ± 0,6	86,3 ± 0,6

- Todos os valores estão expressos em porcentagem.

Tabela 20 – Resultados cos classificadores de cada um dos níveis hierárquicos.

Modelo	Par	Acurácia	AUC	F1-*Score*	Precisão	*Recall*
efn	Lado A x Lado B	85,8 ± 0,5	94,2 ± 0,3	85,2 ± 0,4	84,7 ± 0,5	85,9 ± 0,4
efn	LYN x NÃO LYN	93,4 ± 1,6	96,2 ± 0,4	90,8 ± 2,0	88,0 ± 2,5	95,2 ± 0,7
efn	DEB x NÃO DEB	89,9 ± 0,6	95,8 ± 0,2	87,4 ± 0,6	85,5 ± 0,7	90,2 ± 0,6
efn	MUS x NÃO MUS	94,6 ± 0,3	95,2 ± 0,3	94,0 ± 0,3	93,8 ± 0,4	94,1 ± 0,4
efn	MUC x ADI	88,9 ± 0,7	95,8 ± 0,2	88,9 ± 0,8	89,0 ± 0,7	88,9 ± 0,7
efn	STR x NÃO STR	86,2 ± 0,5	94,2 ± 0,2	84,9 ± 0,4	84,2 ± 0,6	86,1 ± 0,5
efn	NOR x TUM	84,4 ± 0,7	92,2 ± 0,2	84,4 ± 0,7	84,5 ± 0,7	84,4 ± 0,7

- Todos os valores estão expressos em porcentagem.