

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Biociências, Letras e Ciências Exatas- campus de São José do Rio Preto

PEDRO LUCAS MIGUEL

**Integração de Mecanismos de Atenção em Redes Neurais Convolucionais e Híbridas no
Contexto de Imagens Histológicas**

São José do Rio Preto

2026

Pedro Lucas Miguel

**Integração de Mecanismos de Atenção em Redes Neurais Convolucionais e Híbridas no
Contexto de Imagens Histológicas**

Dissertação, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Computação Aplicada

Orientador: Prof^o Dr. Leandro Alves Neves

São José do Rio Preto

2026

M636i	Miguel, Pedro Lucas Integração de Mecanismos de Atenção em Redes Neurais Convolucionais e Híbridas no Contexto de Imagens Histológicas / Pedro Lucas Miguel. -- São José do Rio Preto, 2026 76 p. Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto Orientador: Leandro Alves Neves 1. Inteligência artificial. 2. Visão por computador. 3. Transferência de aprendizagem. 4. Histopatologia. I. Título.
-------	--

IMPACTO POTENCIAL DESTA PESQUISA

O impacto potencial desta pesquisa é a proposição de uma metodologia modular, fundamentada em mecanismos de atenção e regularização por entropia, projetada para ser aplicada em diversas arquiteturas de aprendizado profundo. O objetivo central é o aprimoramento da explicabilidade desses modelos no contexto da análise de imagens histológicas. Paralelamente, o trabalho apresenta um *framework* de avaliação para aferir quantitativamente a qualidade das explicações (via Grad-CAM), fornecendo um meio objetivo para a validação da interpretabilidade dos modelos.

POTENTIAL IMPACT OF THIS RESEARCH

The potential impact of this research is the proposal of a modular methodology, based on attention and entropy regularization mechanisms, designed to be applied in various deep learning architectures. The main objective is to improve the explainability of these models in the context of histological image analysis. At the same time, the work presents an evaluation framework to quantitatively assess the quality of explanations (via Grad-CAM), providing an objective means for validating the interpretability of models.

PEDRO LUCAS MIGUEL

**INTEGRAÇÃO DE MECANISMOS DE ATENÇÃO EM REDES NEURAIS
CONVOLUCIONAIS E HÍBRIDAS NO CONTEXTO DE IMAGENS HISTOLÓGICAS :**

Dissertação apresentado(a) à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Computação Aplicada

Data de defesa: 09/02/2026

BANCA EXAMINADORA

Profº Dr. Leandro Alves Neves
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

Profº Dr. Francisco de Assis Zampirolli
Universidade Federal do ABC

Profº Dr. Geraldo Francisco Donega Zafalon
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

Profº Dr. Sylvio Barbon Junior
University of Trieste, Department of Engineering and Architecture

Profº Dr. Lucas Correia Ribas
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

Dedico este trabalho à minha mãe e a todos os que me auxiliaram na minha jornada.

AGRADECIMENTOS

Agradeço primeiramente à minha mãe, Cira Lucas da Silva, por todo o apoio emocional e financeiro fornecido a mim durante o meu período de graduação e mestrado. Sem esse apoio, este trabalho não seria finalizado.

Agradeço também por todo o suporte fornecido pelo meu psicólogo Pedro Henrique Bazela Desiderio que, ao longo dos últimos 6 anos, vem me auxiliando na conclusão dessa jornada.

Agradeço ao meu orientador, Leandro Alves Neves, por todo o suporte e conhecimento transmitidos a mim durante toda a minha caminhada dentro da universidade.

Por fim, um agradecimento especial aos meus amigos Gabriel, Glauco, Gustavo, Igor, João e Pedro. Agradeço por terem feito parte dessa caminhada e por todo o suporte emocional fornecido.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e pela WZTECH NETWORKS, São José do Rio Preto, São Paulo.

*“Peindre l’amour, peindre la vie, pleurer en couleur. Sur la toile notre amour demeure, se meurt,
je t’aime.”*

Une Vie À T’aimer, Lorien Testard (2025)

RESUMO

Modelos de aprendizado profundo para classificação de imagens histológicas carecem de interpretabilidade. Este trabalho propõe um framework modular, combinando *attention branches* e regularização por entropia (*CAM Fostering*) para aprimorar a explicabilidade do método Grad-CAM. A metodologia foi validada em sete *backbones* (ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer), em cinco *datasets* de imagens histológicas, a partir do uso de métricas quantitativas, sendo elas: complexidade (COM), coerência (CO), queda de confiança (QC) e média DCC (ADCC). Os resultados confirmaram ganhos consistentes no ADCC em seis dos sete modelos, destacando-se o ResNet-50 (63,21%) e o ResNeXt-50 (62,12%). Além disso, observou-se um ganho expressivo na estabilidade das explicações, no ConvNext, por exemplo, a QC média caiu de 19,58% para 5,91%, e modelos híbridos, como o Uniformer, também apresentaram melhorias relevantes, com ADCC de 54,81%. Esses ganhos de interpretabilidade não resultaram em perdas significativas de acurácia, de modo que o desempenho de classificação foi preservado e, em alguns casos, inclusive aprimorado, como no CoatNet-small, que alcançou uma *F1-Score* de 95,14%. Nos demais modelos *backbone*, as variações máximas permaneceram inferiores a 2,2%. Este trabalho oferece, portanto, uma solução capaz de aumentar a transparência do diagnóstico assistido por modelos de aprendizado profundo ao destacar regiões relevantes sem comprometer a acurácia.

Palavras-Chave: attention branches; cam fostering; redes neurais convolucionais; vision transformers; grad-cam; imagens histológicas.

ABSTRACT

Deep learning models for histological image classification lack interpretability. This work proposes a modular framework, combining attention branches and entropy regularization (CAM Fostering), to enhance the explainability of the Grad-CAM method. The methodology was validated across seven backbones (ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNeXt, CoatNet-small, and Uniformer) on five histological image datasets, using quantitative metrics, namely: Complexity (COM), Coherence (CO), Confidence Drop (CD), and Average DCC (ADCC). Results confirmed consistent gains in ADCC in six out of the seven models, with ResNet-50 (63.21%) and ResNeXt-50 (62.12%) standing out. Furthermore, a significant gain in explanation stability was observed; in ConvNeXt, for instance, the average CD dropped from 19.58% to 5.91%, and hybrid models, such as Uniformer, also showed relevant improvements, reaching an ADCC of 54.81%. These interpretability gains did not result in significant accuracy losses; classification performance was preserved and, in some cases, even improved, as seen in CoatNet-small, which achieved an F1-Score of 95.14%. In the other backbone models, maximum variations remained below 2.2%. Therefore, this work offers a solution capable of increasing the transparency of deep learning-assisted diagnosis by highlighting relevant regions without compromising accuracy.

Keywords: attention branches; cam fostering; convolutional neural networks; vision transformers; grad-cam; histological images.

LISTA DE FIGURAS

Figura 1	Amostras pigmentadas pelos compostos H&E referentes a tecidos de câncer mama, em que A e B são amostras benignas e C e D são amostras malignas.	22
Figura 2	Ilustração da arquitetura clássica de uma CNN.	23
Figura 3	Ilustração das características obtidas pelo modelo DenseNet-201, em que: A é a imagem de entrada; B representa as características extraídas de nove filtros da primeira camada convolucional; e C representa as características extraídas de nove filtros da última camada convolucional.	24
Figura 4	Exemplo de <i>max pooling</i> e <i>average pooling</i> (Tamanho do filtro = 2; Passo = 2).	24
Figura 5	Arquitetura de um modelo baseado em <i>transformer</i> .	25
Figura 6	Arquitetura de um modelo tradicional de <i>vision transformers</i> .	27
Figura 7	Diagrama ilustrativo referente a relação entre explicabilidade e acurácia para modelos de aprendizado de máquina.	28
Figura 8	Implementação da camada de GAP em uma CNN para a geração dos mapas de ativação.	29
Figura 9	Aplicação do método Grad-CAM para as classes gato (imagem B) e cachorro (imagem C).	30
Figura 10	Aplicação do método de <i>attention rollout</i> considerando a fusão das matrizes A'_{ij} a partir do cálculo da média e do mínimo.	32
Figura 11	Aplicação do método de <i>attention rollout</i> referente as classes de cachorro e gato.	32
Figura 12	Explicações obtidas pelo modelo XCNN em comparação aos métodos <i>gradiente</i> , <i>guided backprop</i> , <i>deconvolution</i> , <i>deep Taylor decomposition</i> , <i>LRP</i> , <i>PatternNet</i> e <i>PatternAttribute</i> .	39
Figura 13	Explicações obtidas pelo modelo ABN no <i>dataset</i> CelebA.	40
Figura 14	Explicações obtidas pelo modelo proposto por Sarp et al., considerando um caso de pneumonia causado pela COVID-19 e por outras causas, respectivamente.	41
Figura 15	Explicações obtidas pelo método LIME referentes às classes Normal, COVID-19, <i>Malignant Lung Cancer</i> e CAP.	43
Figura 16	Explicações obtidas pelo método SHAP referentes às classes Normal, COVID-19, <i>Malignant Lung Cancer</i> e CAP.	43
Figura 17	Descrição da explicação referente a tomada de decisão, em que, PD é a previsão, GT é o <i>ground truth</i> , R é a razão e A é a ação.	44
Figura 18	Exemplos dos mapas de atenção gerados pela metodologia proposta Dong et al. (2023).	44

Figura 19	Diagrama esquemático do método proposto.	46
Figura 20	Exemplos de imagens histológicas presentes em cada <i>dataset</i>	47
Figura 21	Diagrama esquemático do modelo proposto utilizando camadas convolu- cionais para a composição do <i>attention branch</i>	49
Figura 22	Comparação visual das explicações geradas pelo método Grad-CAM pro- duzidas pelos modelos <i>baseline</i> (esquerda) e proposto (direita), em que: (A) ResNet-50, (B) DenseNet-201, (C) EfficientNet-b0, (D) ResNeXt-50, (E) ConvNeXt (F) CoatNet-small e (G) Uniformer.	61

LISTA DE TABELAS

Tabela 1	– Visão geral de todos os cinco <i>datasets</i>	48
Tabela 2	– Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> UCSB, incluindo CO, COM, QC e ADCC.	55
Tabela 3	– Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> NHL, incluindo CO, COM, QC e ADCC.	55
Tabela 4	– Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> CR, incluindo CO, COM, QC e ADCC.	55
Tabela 5	– Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> LG, incluindo coerência (CO), complexidade (COM), queda de confiança (QC), e ADCC.	56
Tabela 6	– Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> LA, incluindo CO, COM, QC e ADCC.	56
Tabela 7	– Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> UCSB, incluindo CO, COM, QC e ADCC.	58
Tabela 8	– Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> NHL, incluindo CO, COM, QC e ADCC.	58
Tabela 9	– Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> CR, incluindo CO, COM, QC e ADCC.	58
Tabela 10	– Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> LG, incluindo CO, COM, QC e ADCC.	59
Tabela 11	– Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao <i>dataset</i> LA, incluindo CO, COM, QC e ADCC.	59
Tabela 12	– Comparação das médias das métricas de explicabilidade (CO e COM) antes e depois da aplicação da metodologia proposta.	62
Tabela 13	– Comparação das médias das métricas de explicabilidade (QC e ADCC) antes e depois da aplicação da metodologia proposta.	63

Tabela 14 – Média de ACC e *F1-score* (%) de todos os *datasets* para os modelos *baseline* e propostos. 64

LISTA DE ABREVIATURAS E SIGLAS

ABN	Attention Branch Network
ACC	Acurácia
ADCC	Average DCC
AGEMAP	Atlas of Gene Expression in Mouse Ageing Project
CAM	Class Activation Map
CNN	Convolutional Neural Network
CO	Coerência
COM	Complexidade
FN	Falso Negativo
FP	Falso Positivo
GAN	Generative Adversarial Network
GAP	Global Average Pooling
Grad-CAM	Gradient-weighted Class Activation Mapping
GRU	Gated Recurrent Unit
H&E	Hematoxilina e Eosina
K	Key
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
NLP	Natural Language Processing
PR	Precisão
Q	Query
QC	Queda de Confiança
RC	Recall
RNN	Recurrent Neural Network

SHAP	Shapley Additive Explanations
V	Value
ViT	Vision Transformers
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XAI	eXplainable Artificial Intelligence
XCNN	Explainable Convolutional Neural Network

SUMÁRIO

1	INTRODUÇÃO	17
1.1	MOTIVAÇÃO E JUSTIFICATIVAS	18
1.2	OBJETIVOS	20
1.3	ORGANIZAÇÃO DO TRABALHO	20
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	IMAGEM DIGITAL	21
2.2	IMAGENS HISTOLÓGICAS	21
2.3	REDES NEURAIS CONVOLUCIONAIS	22
2.4	MODELOS BASEADOS EM <i>TRANSFORMERS</i>	23
2.4.1	<i>Vision Transformers</i>	26
2.5	INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL	27
2.5.1	<i>Class Activation Map</i>	28
2.5.2	<i>Gradient-weighted Class Activation Mapping</i>	30
2.5.3	<i>Attention Rollout</i>	31
2.6	MÉTRICAS PARA A AVALIAÇÃO DAS EXPLICAÇÕES	33
2.6.1	Coerência (CO)	33
2.6.2	Complexidade (COM)	33
2.6.3	Queda de Confiança (QC)	34
2.6.4	<i>Average DCC</i> (ADCC)	34
2.7	MÉTRICAS PARA A AVALIAÇÃO DE DESEMPENHO DE MODELOS DE APRENDIZADO PROFUNDO	34
2.7.1	Acurácia (ACC)	35
2.7.2	Precisão (PR)	35
2.7.3	Recall (RC)	36
2.7.4	<i>F1-score</i>	36
3	TRABALHOS RELACIONADOS	38
3.1	REDES NEURAIS CONVOLUCIONAIS EXPLICÁVEIS	38
3.2	EXPLICABILIDADE EM MODELOS VIT	42
3.3	CONSIDERAÇÕES SOBRE OS TRABALHOS CORRELATOS	45
4	METODOLOGIA	46
4.1	DEFINIÇÃO E PARTICIONAMENTO DOS <i>DATASETS</i> H&E	46
4.2	DEFINIÇÃO E TREINAMENTO DO MODELO PROPOSTO: ABN COM <i>CAM FOSTERING</i>	48

4.2.1	<i>Feature Extractor</i>	48
4.2.2	<i>Attention Branch</i>	49
4.2.3	Mecanismo de Atenção	50
4.2.4	<i>Perception Branch</i>	50
4.2.5	<i>CAM Fostering</i>	51
4.2.6	Treinamento dos Modelos	52
4.3	DEFINIÇÃO DO PROCESSO DE AVALIAÇÃO DAS EXPLICAÇÕES . .	52
4.3.1	Pacotes de Software e Ambiente de Execução	53
5	RESULTADOS E DISCUSSÃO	54
5.1	MODELOS <i>BACKBONE</i> ORIGINAIS	54
5.2	AVALIAÇÃO DOS MODELOS PROPOSTOS	57
5.2.1	Análise Qualitativa das Explicações	60
5.2.2	Resumo Comparativo entre os Modelos Propostos e <i>Baseline</i>	62
5.3	VISÃO GERAL SOBRE OS DESEMPENHOS DE CLASSIFICAÇÃO . . .	64
6	CONCLUSÃO	65
6.1	PUBLICAÇÕES	66
	REFERÊNCIAS	67
	DADOS CURRICULARES	77

1 INTRODUÇÃO

Nos últimos anos, modelos de aprendizado profundo baseados em redes neurais convolucionais (CNN) e *vision transformers* (ViT) (Vaswani et al., 2023) têm alcançado resultados considerados estado da arte em diversas tarefas, incluindo a classificação de imagens digitais (Liu; Wang; Yue, 2024; Dwivedi; Dutta; Pandey, 2024). Esses avanços impulsionaram o desenvolvimento de sistemas computacionais capazes de auxiliar em áreas críticas da sociedade, como a medicina, na qual a análise automatizada de imagens histológicas se destaca como uma ferramenta essencial nos processos de diagnóstico e prognóstico de doenças (Roberto et al., 2024; Tenguam et al., 2024; Rozendo et al., 2022; Höhn et al., 2021; Shihabuddin; K., 2023; Majumdar; Pramanik; Sarkar, 2023).

A análise histopatológica é de extrema importância no processo de diagnóstico de doenças que afetam tecidos biológicos. Esse processo envolve: a coleta de amostras de tecidos; pigmentação das amostras por meio de compostos químicos como hematoxilina e eosina (H&E) e análise em microscópio para identificar possíveis anomalias morfológicas (Fischer et al., 2008). Embora indispensável para a prática clínica, esse tipo de atividade demanda elevado investimento de tempo e apresenta uma variabilidade considerável, tendo em vista que a mesma depende da competência e do critério individual dos especialistas (Dobbs et al., 2015). Neste contexto, o desenvolvimento de sistemas computacionais baseados em CNNs e ViTs pode auxiliar consideravelmente o processo de análise histológica, de forma que seja possível tornar o processo ainda mais ágil e preciso (Oliveira et al., 2024; Pan et al., 2025; Li et al., 2025; Arrieta Legorburu et al., 2025; Durand R. et al., 2025; Li et al., 2022a).

No entanto, apesar de seu alto desempenho de classificação, modelos CNN e ViT apresentam uma interpretabilidade limitada, resultando em uma menor transparência de seus processos internos de tomada de decisão, dificultando sua adoção na prática clínica (Szandała, 2023; Chau; Rahman; Debnath, 2025). Essa falta de interpretabilidade gera uma série de questões quanto à confiabilidade desses sistemas, o que é especialmente importante na área médica (Xu et al., 2019). Em resposta a esse desafio, o campo da inteligência artificial explicável (XAI) cresceu rapidamente, concentrando-se em técnicas que tornam as classificações de modelos mais compreensíveis e confiáveis (Rozendo et al., 2024; Longo et al., 2024; Martinez et al., 2024). Dentre as técnicas XAI, o *gradient-weighted class activation mapping* (Grad-CAM) (Selvaraju et al., 2017) é uma das mais utilizadas com modelos convolucionais. Essa técnica se baseia na utilização dos valores de gradiente das camadas convolucionais para a geração de mapas de calor, indicando as regiões da imagem de entrada que mais influenciaram na classificação final do modelo. Já para modelos ViT, o *attention-rollout* (Abnar; Zuidema, 2020) é frequentemente utilizado para combinar valores de atenção de várias camadas e *attention-heads* em uma visualização unificada. Essas estratégias permitem a visualização do foco e da relevância do modelo, servindo como uma ponte entre os resultados do modelo e o raciocínio humano.

Além do desenvolvimento de técnicas para geração de explicações, diferentes arquiteturas de redes neurais foram projetadas especialmente para aprimorar a explicabilidade intrínseca de modelos (Iglesias; Menendez; Talavera, 2025; Ayaz et al., 2024; Tsukahara et al., 2021; Miguel et al., 2024). A *attention branch network* (ABN) (Fukui et al., 2019) aprimora a explicabilidade de *backbones* convolucionais mediante um bloco *attention branch*. Este bloco tem a função de auxiliar a rede na identificação de características mais relevantes da imagem de entrada durante o treinamento, aprimorando a informatividade dos mapas de calor gerados por técnicas como o Grad-CAM. Da mesma forma, a *explainable convolutional neural network* (XCNN) (Tavanaei, 2020) aproveita uma estrutura baseada em codificadores e decodificadores para gerar e refinar mapas de atenção, apoiada por um discriminador que incentiva a fidelidade e a relevância nas explicações aprendidas. Mais recentemente, estratégias como o *CAM Fostering* passaram a empregar a entropia como forma de regular a qualidade das explicações baseadas em mapas de ativação (Schöttl, 2022). A entropia, nesse contexto, indica o grau de dispersão ou incerteza presente na distribuição espacial das ativações: mapas com baixa entropia tendem a concentrar a atenção em poucas regiões ou, inversamente, a distribuí-la de forma quase uniforme, ambos os casos resultando em explicações pouco informativas (Jain; Wallace, 2019; Bylinskii et al., 2017). Ao penalizar mapas de baixa entropia, essa técnica incentiva o modelo a produzir ativações mais equilibradas, promovendo diversidade espacial e ressaltando regiões verdadeiramente relevantes do ponto de vista semântico. Essa forma de regularização tem se mostrado promissora não apenas para aprimorar a interpretabilidade, mas também para favorecer a generalização dos modelos de aprendizagem profunda.

1.1 MOTIVAÇÃO E JUSTIFICATIVAS

Apesar dos avanços significativos no uso de redes CNNs e ViTs em tarefas de análise de imagens médicas, ainda persistem lacunas relevantes na literatura. Um dos principais desafios diz respeito à escassez de estratégias de explicabilidade desenvolvidas e validadas especificamente para imagens histológicas, que possuem características visuais muito distintas de imagens naturais. Essas imagens são marcadas por texturas altamente heterogêneas, estruturas organizadas em múltiplas escalas e variações morfológicas sutis que demandam métodos interpretativos adaptados ao domínio biomédico (Dobbs et al., 2015; Kashefi et al., 2023). Outro aspecto importante envolve a avaliação da maioria das técnicas XAI, que se apoiam majoritariamente em métricas qualitativas, especialmente na análise visual subjetiva das regiões identificadas como relevantes pelos modelos (Ioannidis; Maniadis, 2023). Embora análises qualitativas permitam uma avaliação mais imediata dos resultados, esse tipo de abordagem compromete a reprodutibilidade e dificulta a comparação sistemática entre métodos. Além disso, a ausência de métricas quantitativas padronizadas para avaliar a qualidade das explicações pode limitar o avanço metodológico da área.

No que se refere aos ViTs, embora esses modelos tenham se mostrado altamente eficazes

em tarefas de classificação e segmentação de imagens médicas (Liu; Hu; Chen, 2023; Islam et al., 2024; Mahmud Kabir; Imamul Hassan Bhuiyan, 2025), seu potencial explicativo ainda não foi explorado de forma sistemática. Devido à sua arquitetura baseada em mecanismos de *self-attention* (Vaswani et al., 2023), os ViTs possuem uma capacidade inerente de capturar relações espaciais globais, o que os torna particularmente promissores em cenários em que os padrões relevantes estão distribuídos ou são discretos, como ocorre frequentemente em amostras histopatológicas. No entanto, ainda são escassas as abordagens que buscam aproveitar essas propriedades para produzir explicações mais ricas, informativas e confiáveis.

No que diz respeito aos avanços proporcionados pelas arquiteturas ABN (Fukui et al., 2019), que introduziram blocos de atenção supervisionada para guiar os modelos a destacar regiões críticas durante o treinamento, melhorando a qualidade dos mapas de atenção gerados por técnicas como o Grad-CAM, e pelas estratégias como o *CAM Fostering*, que utilizam regularização baseada na entropia para promover explicações espacialmente diversificadas e semanticamente coerentes, evitando tanto a dispersão quanto a superconcentração da atenção (Schöttl, 2022), ainda persistem desafios relevantes. Em particular, a integração dessas abordagens para ampliar a qualidade explicativa dos modelos sem comprometer sua acurácia classificatória, aspecto fundamental para a adoção clínica desses sistemas, permanece pouco explorada.

A partir dessas constatações, duas questões centrais merecem destaque:

- **Q1.** A integração do bloco de atenção do modelo ABN com regularização baseada em entropia (*CAM Fostering*) melhora a capacidade explicativa de arquiteturas convolucionais e híbridas aplicadas a imagens histopatológicas H&E?
- **Q2.** Esse tipo de integração mantém a capacidade de explicação e classificação dos modelos convolucionais e híbridos em diferentes contextos de amostras H&E?

Considerando os direcionamentos delineados e as lacunas identificadas, as hipóteses que orientam esta investigação foram as seguintes:

- **H1.1:** A combinação de atenção supervisionada com regularização por entropia resultará em explicações significativamente mais informativas que as versões *baseline* dos mesmos *backbones*.
- **H1.2:** A regularização por entropia (*CAM Fostering*), quando aplicada simultaneamente com mecanismos de atenção supervisionada, aprimorará quantitativamente a capacidade explicativa dos modelos propostos em relação aos *baselines*.
- **H2.1:** Essa integração manterá taxas de classificação estáveis ou superiores mesmo em conjuntos de dados com número reduzido de amostras, demonstrando capacidade de generalização.

Por fim, a motivação central deste trabalho focou na necessidade de tornar modelos de aprendizado profundo mais interpretáveis, confiáveis e úteis às especificidades da histopatologia

digital. Ao investigar a eficácia de uma arquitetura explicável que combina supervisão de atenção com regularização por entropia, este trabalho contribui com o desenvolvimento de soluções computacionais mais transparentes e aplicáveis à prática clínica.

1.2 OBJETIVOS

Neste trabalho, a arquitetura ABN foi aprimorada por meio da incorporação da estratégia *CAM Fostering*, com o objetivo de melhorar a interpretabilidade e a qualidade das regiões discriminativas geradas por modelos *backbone* na análise de imagens histológicas. Os objetivos específicos foram:

- Desenvolver uma arquitetura modular explicável, combinando mecanismos de atenção e regularização baseada em *CAM Fostering*, compatível com modelos convolucionais e híbridos, visando aprimorar a qualidade e a relevância das explicações na classificação de imagens histológicas;
- Avaliar as explicações geradas pelo método proposto a partir de um *framework* de avaliação baseado em métricas quantitativas, permitindo, assim, aferir objetivamente a qualidade das explicações visuais geradas por modelos de aprendizado profundo;
- Identificar o impacto das modificações propostas em relação às taxas de classificação, comparando os resultados obtidos pelos modelos propostos em relação aos modelos *baseline*;

1.3 ORGANIZAÇÃO DO TRABALHO

No Capítulo 2 é apresentada a fundamentação teórica que possibilitou o desenvolvimento deste trabalho. Já no Capítulo 3 são apresentados os trabalhos relacionados ao método aqui proposto. No Capítulo 4 é apresentada a metodologia utilizada para o desenvolvimento deste trabalho. No Capítulo 5, são apresentados os resultados obtidos a partir da aplicação da metodologia, apresentando comparações diretas entre o método proposto e os modelos *baseline* no quesito de qualidade de explicações no contexto de imagens histológicas. Por fim, no Capítulo 6, é apresentada a conclusão deste trabalho, bem como projetos futuros seguindo a mesma linha.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os fundamentos utilizados para o desenvolvimento deste trabalho, com informações sobre conceitos de imagens digitais, imagens histológicas, redes neurais convolucionais, inteligência artificial explicável, métricas para avaliação de explicações e, por fim, modelos baseados na arquitetura de *transformers*, em especial modelos baseados em *vision transformers*.

2.1 IMAGEM DIGITAL

Uma imagem pode ser representada por uma função $I(x, y)$, em que cada coordenada espacial (x, y) está associada a um valor de luminância. Quando tanto a função I quanto as coordenadas (x, y) pertencem a um espaço discreto, a imagem é caracterizada como uma imagem digital, cujos elementos (*pixels*) são organizados em linhas e colunas dentro de uma região finita (Gonzalez; Woods, 2006).

A função $I(x, y)$ pode ser descrita em termos de iluminação e refletância. No caso de uma imagem digital obtida por um dispositivo de captura, como uma câmera ou um microscópio, ela pode ser definida por meio da seguinte função pontual:

$$I(x, y) = i(x, y)r(x, y), \quad (1)$$

em que i representa a função de iluminação e r a função de refletância, ambas dependentes do ponto (x, y) . O produto dessas funções, $i(x, y) \cdot r(x, y)$, determina a luminância em um dado ponto (x, y) .

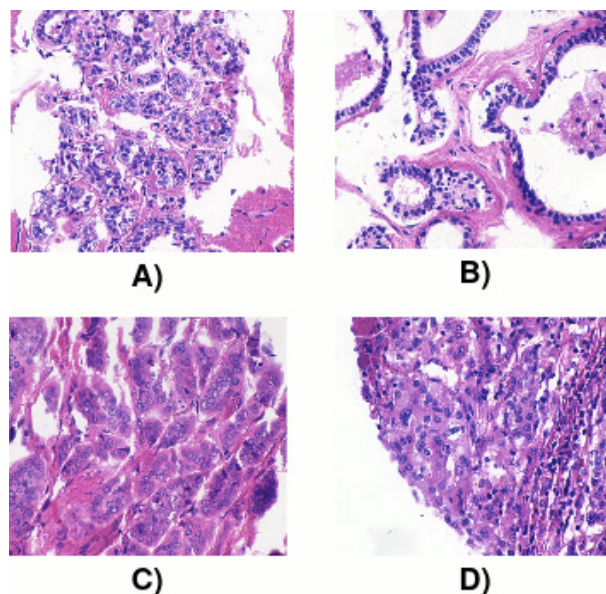
2.2 IMAGENS HISTOLÓGICAS

A análise de imagens histológicas constitui uma etapa fundamental nos processos de diagnóstico e prognóstico de diversas doenças e condições relacionadas aos tecidos humanos. Essas imagens são geralmente obtidas por meio de microscópios ou dispositivos especializados, como *scanners* (Filipiak et al., 2014).

A aquisição das imagens digitais de amostras de tecido envolve uma sequência de etapas, incluindo a coleta da amostra, fixação, processamento, inclusão, pigmentação e análise por equipamentos especializados (Gurina; Simms, 2023). Dentre essas etapas, a pigmentação assume papel crucial, pois possibilita uma melhor visualização e compreensão da estrutura tecidual. Entre as técnicas de pigmentação, destaca-se a utilização dos compostos químicos hematoxilina e eosina (H&E). A hematoxilina é responsável pela coloração dos ácidos nucleicos presentes no tecido, conferindo-lhes uma tonalidade escura que varia entre o azul e o lilás. Por sua vez, a

eosina pigmenta as proteínas, apresentando-as em tons rosados (Fischer et al., 2008). Na Figura 1 são ilustra algumas amostras de tecido coradas pelos compostos H&E.

Figura 1 – Amostras pigmentadas pelos compostos H&E referentes a tecidos de câncer mama, em que A e B são amostras benignas e C e D são amostras malignas.



Fonte: Universidade da Califórnia, Santa Bárbara (Gelasca et al., 2008).

Esse procedimento não apenas permite que especialistas realizem uma investigação mais precisa das amostras de tecido, mas também viabiliza a implementação de sistemas computacionais avançados para análise, diagnóstico e prognóstico de doenças. Destacam-se, nesse contexto, os sistemas baseados em redes neurais convolucionais e *vision transformers* (Silva et al., 2022; Yang et al., 2024; Kanwal et al., 2024; Hörst et al., 2024).

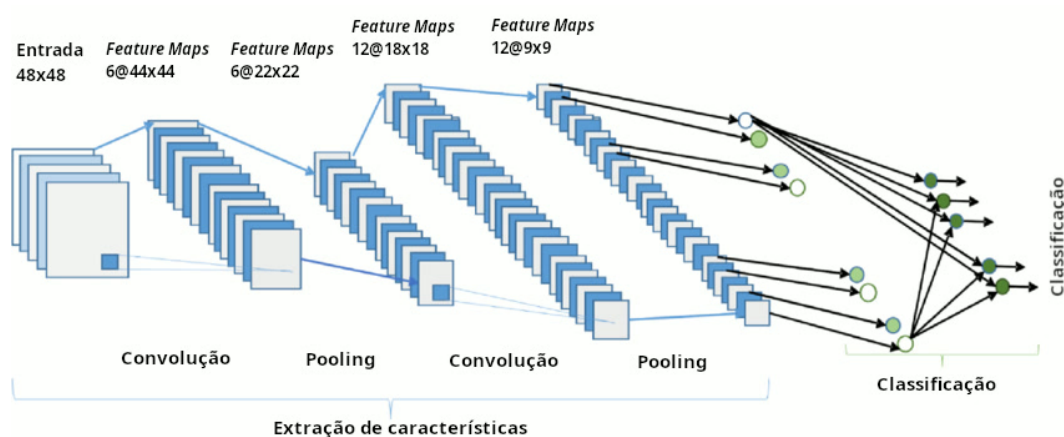
2.3 REDES NEURAIAS CONVOLUCIONAIS

Redes Neurais Convolucionais, do inglês *Convolutional Neural Networks* (CNNs), constituem ferramentas robustas para a análise de dados estruturados (LeCun; Bengio; Hinton, 2015). A topologia de uma CNN é fundamentada em três conceitos principais: campos receptivos locais, pesos compartilhados e amostragem espacial ou temporal (LeCun et al., 1998). Tais conceitos se refletem nas diferentes camadas que compõem uma CNN, cada uma responsável pela extração de características específicas dos dados de entrada.

Sistemas computacionais baseados em CNNs têm ganhado crescente destaque na classificação e no reconhecimento de padrões em imagens digitais, especialmente na área médica (Höhn et al., 2021; Shihabuddin; K., 2023; Majumdar; Pramanik; Sarkar, 2023). Essa popularidade decorre, em grande parte, dos resultados expressivos apresentados por arquiteturas testadas na competição ILSVRC2012 (Russakovsky et al., 2015), com destaque para o modelo AlexNet (Krizhevsky; Sutskever; Hinton, 2012).

A arquitetura de uma CNN é composta por camadas convolucionais intercaladas por camadas de *pooling*, que, por sua vez, são seguidas por camadas totalmente conectadas (Zhou; Greenspan; Shen, 2017), conforme ilustrado na Figura 2. As camadas convolucionais são responsáveis pela extração de características dos dados de entrada, armazenando essas informações nos chamados mapas de características (*feature maps*). O tipo de característica extraída pelas camadas convolucionais varia conforme sua posição na arquitetura (Santos; Ponti, 2019). Camadas localizadas no início da rede extraem características locais dos dados, como cantos, bordas e texturas em imagens digitais. Já as camadas posicionadas mais ao final são capazes de capturar características mais complexas. Na Figura 3 é ilustrada as diferenças nas características extraídas conforme a localização das camadas convolucionais na rede.

Figura 2 – Ilustração da arquitetura clássica de uma CNN.



Fonte: Adaptado de Alom et al. (2019).

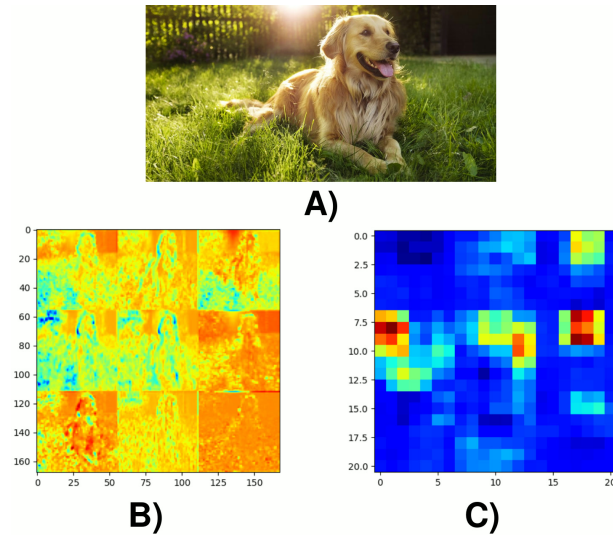
As camadas de *pooling* são responsáveis pela redução dos mapas de características gerados pelas camadas convolucionais (Affonso et al., 2017), tendo como objetivo principal garantir que a rede foque nos padrões mais relevantes. Esse processo promove a robustez da rede frente a pequenas variações nos padrões aprendidos (Namatêvs, 2017).

Para realizar o *pooling*, é necessário definir dois hiperparâmetros: o tamanho da janela (*filter*) e o deslocamento dessa janela (*stride*). Conforme a janela desliza sobre o mapa de características, as operações mais comuns são o *max pooling* e o *average pooling* (GEEKS FOR GEEKS, 2025). O *max pooling* seleciona o maior valor dentro da janela para compor a matriz resultante, enquanto o *average pooling* calcula a média dos valores presentes na janela para definir o valor de saída. Na Figura 4 é ilustrado como cada um destes métodos atuam individualmente.

2.4 MODELOS BASEADOS EM TRANSFORMERS

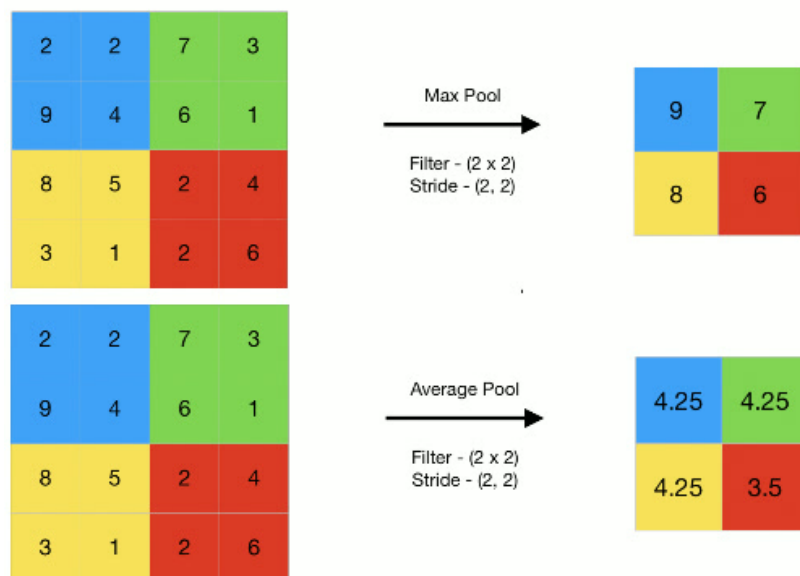
Nos últimos anos, modelos baseados em *transformers* vêm ganhando crescente destaque em diversas áreas, especialmente no contexto de processamento de linguagem natural, do inglês *natural language processing* (NLP) (Huang et al., 2024). Essa popularidade deve-se ao desenvolvimento de modelos altamente performáticos e de fácil aplicação, como BERT (Devlin

Figura 3 – Ilustração das características obtidas pelo modelo DenseNet-201, em que: A é a imagem de entrada; B representa as características extraídas de nove filtros da primeira camada convolucional; e C representa as características extraídas de nove filtros da última camada convolucional.



Fonte: Elaborado pelo autor.

Figura 4 – Exemplo de *max pooling* e *average pooling* (Tamanho do filtro = 2; Passo = 2).



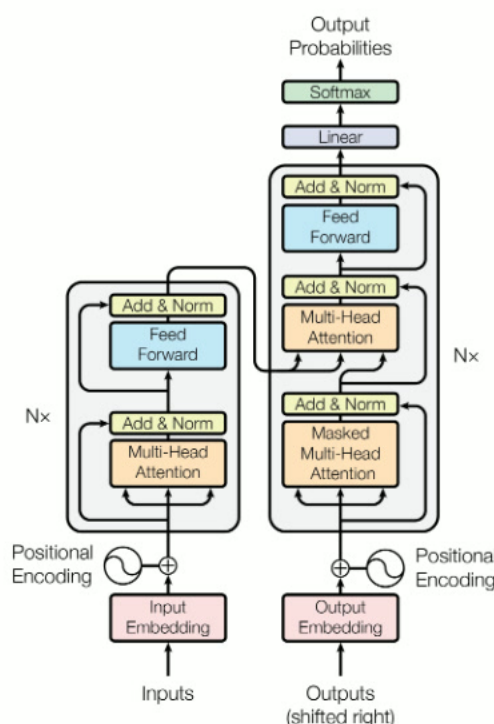
Fonte: Retirado de Geeks For Geeks (2025).

et al., 2019), Llama 2 (Touvron et al., 2023) e GPT (Brown et al., 2020; Radford; Narasimhan, 2018; Radford et al., 2019).

Antes da popularização dos modelos baseados em *transformers*, as redes neurais recorrentes, do inglês *recurrent neural networks* (RNNs), apresentavam os melhores resultados na análise de dados temporais, no processamento de linguagem natural e no reconhecimento de fala (Schmidt, 2019). Esses modelos combinam camadas convolucionais com conexões recorrentes, o que permite manter um estado interno capaz de capturar dependências temporais em sequências de dados. Entretanto, devido à natureza convolucional das RNNs, problemas como a dificuldade em captar dependências temporais de longo prazo e o problema do gradiente de fuga (Hochreiter, 1998) motivaram o desenvolvimento de arquiteturas que dispensassem o uso de convolução, como os *transformers*.

Uma arquitetura tradicional de *transformers* é composta por cinco componentes principais: codificação posicional, mecanismos de *self-attention*, *multi-head attention*, conexões residuais e *feed forward networks* (Vaswani et al., 2023). Na Figura 5 é apresentado um modelo de rede neural baseado na arquitetura de *transformers*.

Figura 5 – Arquitetura de um modelo baseado em *transformer*.



Fonte: Retirado de Vaswani et al. (2023).

A codificação posicional é responsável por fornecer ao modelo informações sobre a ordem dos *tokens*, que podem corresponder a palavras ou até mesmo letras. Esse processo é frequentemente realizado por meio de funções sinusoidais, o que permite à arquitetura discernir a sequência dos *tokens* e capturar a dependência temporal entre eles.

O mecanismo de *self-attention* constitui uma parte essencial da arquitetura de *transformers*. Diferentemente das RNNs e CNNs tradicionais, que processam os dados sequencialmente, o *transformer* utiliza o *self-attention* para calcular a importância de cada *token* considerando o contexto global da sequência. Para cada *token*, são associados três vetores: *query* (Q), *key* (K) e *value* (V). Esses vetores são empregados para calcular os chamados *attention scores* entre pares de *tokens*, possibilitando a agregação eficiente da informação de toda a sequência e, consequentemente, uma captura mais efetiva das dependências temporais em sequências longas.

O mecanismo de *multi-head self-attention* possibilita a paralelização eficiente da arquitetura de *transformers*. Para uma sequência de entrada qualquer, essa é processada por múltiplas *attention heads*, cada uma focada em diferentes aspectos dos dados de entrada. Assim, o modelo consegue extrair informações intrínsecas, bem como as relações entre os elementos da sequência. As saídas dessas múltiplas *attention heads* são concatenadas e transformadas linearmente para gerar a representação final do processo de *self-attention*.

As conexões residuais têm o papel de facilitar o treinamento da arquitetura, especialmente em redes profundas, nas quais aumentam as chances de problemas durante o processo de aprendizagem, situação semelhante à observada em modelos RNN. Essas conexões promovem uma melhor propagação do gradiente, mitigando problemas como o gradiente de fuga e, consequentemente, facilitam o treinamento do modelo (He et al., 2016).

Por fim, as *feed forward networks* desempenham papel fundamental na arquitetura de *transformers*, processando os dados obtidos pelos blocos de atenção. Essas redes são constituídas por duas camadas totalmente conectadas, intercaladas por uma função de ativação ReLU. Por meio dessas camadas, o modelo captura relações intrínsecas entre os *tokens* e extrai características de mais alto nível da sequência de entrada.

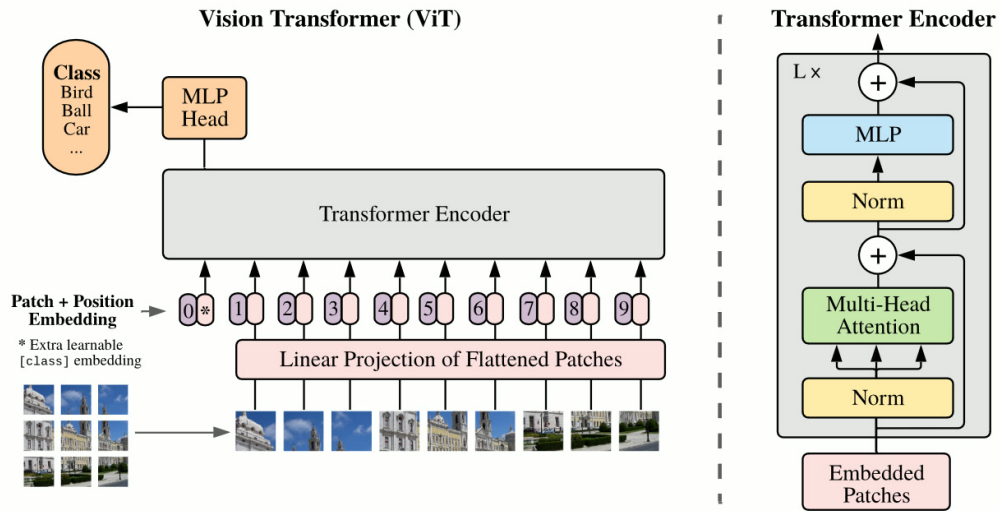
Com a popularização da arquitetura de *transformers*, diversas áreas passaram a explorar sua aplicação em contextos específicos, destacando-se, sobretudo, o processamento de visão computacional, por meio dos chamados *vision transformers*.

2.4.1 Vision Transformers

A aplicação da arquitetura de *transformers* no campo da visão computacional tem proporcionado resultados expressivos em diversas áreas, especialmente na medicina, evidenciando avanços que ainda não haviam sido alcançados por modelos convolucionais tradicionais (Baykal Kablan; Ayas, 2024; Gocer, 2024; Bi et al., 2023).

A implementação da arquitetura de *vision transformers* (ViT) foi viabilizada a partir do processo denominado *patch and positional embedding* (Dosovitskiy et al., 2021). Esse procedimento consiste na divisão da imagem de entrada em pequenos *patches* de tamanho 16×16 , sendo que cada *patch* é convertido em uma representação vetorial, a qual é posteriormente submetida a uma projeção linear, gerando um *token* único para cada *patch*. Por fim, cada *token* recebe um vetor especial que informa ao modelo a relação posicional correspondente. Na Figura 6 é apresentado uma visão geral da arquitetura ViT tradicional.

Figura 6 – Arquitetura de um modelo tradicional de *vision transformers*.



Fonte: Retirado de Dosovitskiy et al. (2021).

Uma das principais vantagens dos modelos de *vision transformers* em relação aos modelos convolucionais tradicionais é a capacidade de interpretar uma imagem considerando seu contexto global. Essa característica decorre da arquitetura de *transformers*, que relaciona a importância de cada *token* em relação aos demais por meio do mecanismo de *self-attention*.

Esse aspecto também contribui para a explicabilidade dos modelos de *vision transformers*, uma vez que a extração dos valores de *self-attention* possibilita a criação de mapas de calor que indicam a relevância de cada região dentro de um determinado *patch*. Consequentemente, a concatenação desses valores gera um mapa de calor único, evidenciando as regiões mais importantes da imagem para a classificação final (Leem; Seo, 2024).

2.5 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Com a popularização inicial dos modelos convolucionais para classificação de imagens digitais, arquiteturas progressivamente mais complexas foram desenvolvidas para superar as limitações das primeiras CNNs (Li et al., 2022b). Nesse cenário, destacam-se as arquiteturas ResNet e DenseNet, que, por meio do uso de conexões residuais, permitiram a criação de modelos mais profundos e com melhor acurácia (He et al., 2016; Huang et al., 2017).

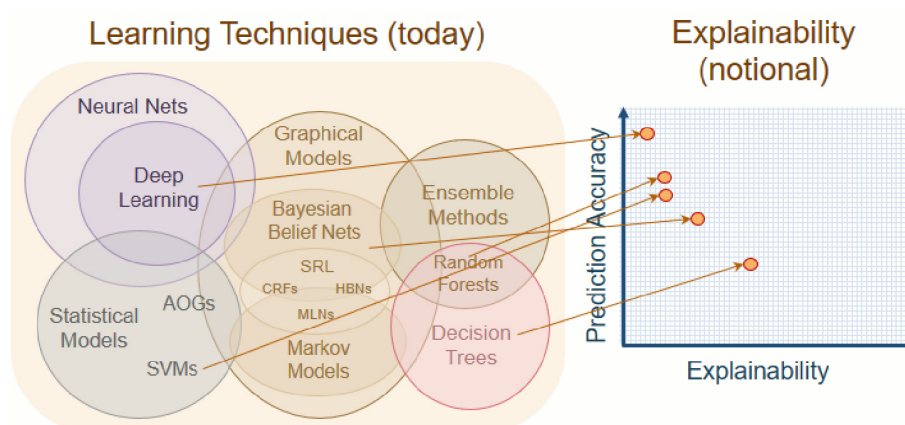
Contudo, esse avanço tornou os modelos significativamente mais complexos, dificultando a compreensão dos processos internos responsáveis pela classificação. Por essa razão, tais arquiteturas passaram a ser referidas na literatura como “caixas pretas” (Szandała, 2023), gerando desconfiância quanto aos parâmetros e características efetivamente relevantes para a decisão final. Essa opacidade limita a aplicação desses sistemas em áreas críticas, como a medicina (Xu et al., 2019).

Nesse contexto, as técnicas de inteligência artificial explicável, do inglês *eXplainable Ar-*

tificial Intelligence (XAI), têm ganhado crescente relevância na última década (Saeed; Omlin, 2023), especialmente no campo das imagens digitais (Kamakshi; Krishnan, 2023).

Um dos maiores desafios no desenvolvimento de métodos de XAI é encontrar um equilíbrio entre acurácia e explicabilidade, uma vez que, frequentemente, o aumento da interpretabilidade dos modelos acarreta uma redução no desempenho preditivo (Bell et al., 2022). Essa relação está ilustrada na Figura 7.

Figura 7 – Diagrama ilustrativo referente a relação entre explicabilidade e acurácia para modelos de aprendizado de máquina.



Fonte: Retirado de DARPA (2024).

Uma das principais vantagens dos modelos de *vision transformers* em relação aos modelos convolucionais tradicionais é a capacidade de interpretar uma imagem considerando seu contexto global. Essa característica decorre da arquitetura de *transformers*, que relaciona a importância de cada *token* em relação aos demais por meio do mecanismo de *self-attention*.

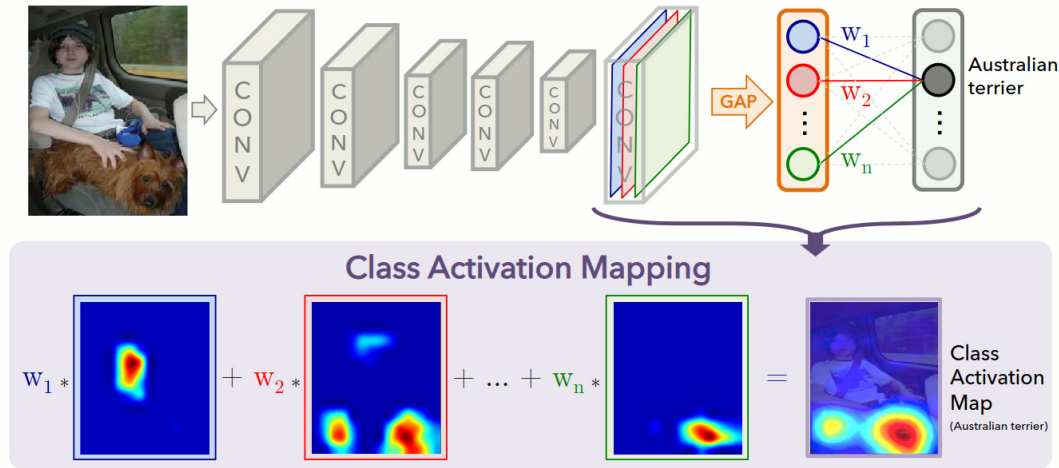
Esse aspecto também contribui para a explicabilidade dos modelos de *vision transformers*, uma vez que a extração dos valores de *self-attention* possibilita a criação de mapas de calor que indicam a relevância de cada região dentro de um determinado *patch*. Consequentemente, a concatenação desses valores gera um mapa de calor único, evidenciando as regiões mais importantes da imagem para a classificação final (Leem; Seo, 2024).

2.5.1 Class Activation Map

O *class activation map* (CAM) foi o primeiro método baseado em mapas de ativação utilizado para a geração de explicações (Zhou et al., 2016). Esse método explora a capacidade intrínseca das camadas convolucionais em uma CNN de localizar objetos em uma imagem (Zhou et al., 2015). Entretanto, essa capacidade tende a ser perdida durante o treinamento caso o modelo utilize uma camada totalmente conectada para a classificação final. Para contornar essa limitação, o método CAM propõe substituir a camada totalmente conectada por uma camada de *global average pooling* (GAP), que atua não apenas como um mecanismo regularizador, mas também

permite que as camadas convolucionais mantenham sua habilidade de localizar objetos na imagem. Na Figura 8 é ilustrada a implementação da camada de GAP em um modelo específico.

Figura 8 – Implementação da camada de GAP em uma CNN para a geração dos mapas de ativação.



Fonte: Retirado de Zhou et al. (2016).

Vale ressaltar que a substituição da camada totalmente conectada por uma camada de GAP exige o retreinamento do modelo, o que torna a implementação dessa abordagem bastante dispendiosa, especialmente em redes de alta complexidade.

Para a geração das explicações com base nesse método, considera-se a função $f_k(x, y)$, na qual k representa uma unidade de ativação da última camada convolucional de uma CNN qualquer, localizada na posição espacial (x, y) . Assim, o processo de aplicação do *global average pooling* sobre essa camada é descrito pela seguinte equação:

$$F_k = \sum_{x,y} f_k(x, y). \quad (2)$$

Assim, para obter os mapas de ativação referentes a uma classe c , é necessário primeiro calcular a pontuação S associada a essa classe, definida pela seguinte equação:

$$S_c = \sum_k w_k^c F_k. \quad (3)$$

em que, S_c é a pontuação da classe c escolhida e w representa o peso correspondente à unidade k da última camada convolucional para essa classe. Dessa forma, é possível ponderar os valores de ativação F_k da última camada convolucional em relação à pontuação S_c , resultando na seguinte equação:

$$\begin{aligned}
S_c &= \sum_k w_k^c \sum_{x,y} f_k(x,y) \\
&= \sum_{x,y} \sum_k w_k^c f_k(x,y).
\end{aligned} \tag{4}$$

Em seguida, define-se M_c como o mapa de ativação referente à classe c , em que cada elemento espacial desse mapa pode ser representado pela Equação 5, indicando a importância de cada ativação na coordenada espacial (x, y) .

$$M_c(x, y) = \sum_k w_k^c f_k(x, y). \tag{5}$$

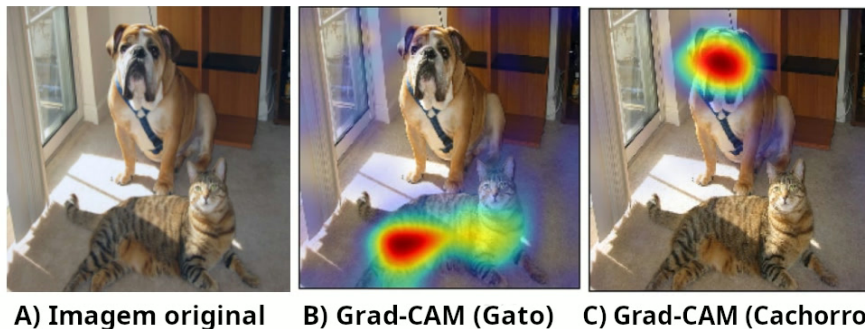
Por fim, o mapa de ativação M_c deve ser submetido a um processo de aumento de resolução, a fim de viabilizar sua sobreposição à imagem original, proporcionando uma visualização das regiões mais relevantes para a classificação na classe c .

2.5.2 Gradient-weighted Class Activation Mapping

O *Gradient-weighted Class Activation Mapping* (Grad-CAM) é uma variação do método CAM que utiliza uma abordagem que dispensa modificações nos modelos de CNN para a geração dos mapas de ativação (Selvaraju et al., 2017). Essa característica facilita a implementação e elimina a necessidade de retreinamento do modelo, constituindo as principais vantagens desta técnica.

Diferentemente do método CAM, que emprega as ativações e pesos da última camada convolucional para gerar os mapas de ativação, o Grad-CAM utiliza não apenas os valores de ativação dessa camada, mas também os gradientes calculados em relação a uma classe c . Esses gradientes são geralmente obtidos durante o processo de *backpropagation* no treinamento da rede e indicam a importância de cada neurônio da última camada convolucional para a classificação da classe c . Na Figura 9 é ilustrada a aplicação desse método.

Figura 9 – Aplicação do método Grad-CAM para as classes gato (imagem B) e cachorro (imagem C).



Fonte: Adaptado de Selvaraju et al. (2017).

Para calcular o Grad-CAM de uma determinada classe, é necessário inicialmente obter os gradientes da pontuação da classe c (y^c) em relação ao mapa de ativações (A^k) da última camada convolucional. Esses gradientes são submetidos a um processo de *global average pooling* para que seja possível determinar a importância de cada neurônio da camada. Esse procedimento está formalmente descrito pela equação abaixo:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \quad (6)$$

em que α representa o valor de importância de cada neurônio k da última camada convolucional em relação à classe c . Com esses valores de importância, aplica-se uma combinação ponderada dos pesos α sobre os valores de ativação da camada A , seguida por uma operação de ReLU, conforme apresentado na equação abaixo:

$$L_{Grad-Cam}^C = ReLU \left(\sum_k \alpha_k^c A^k \right). \quad (7)$$

em que $L_{Grad-Cam}$ representa o mapa de ativação referente à classe c . De forma análoga ao método CAM, esse mapa de ativação é posteriormente submetido a um processo de aumento de resolução, permitindo sua sobreposição à imagem original.

2.5.3 Attention Rollout

Embora as técnicas apresentadas nas seções anteriores apresentem bons resultados no contexto de CNNs, elas não são diretamente aplicáveis a modelos ViT, devido às significativas diferenças arquiteturais entre esses modelos. Assim, diferentes abordagens foram desenvolvidas para viabilizar a geração de explicações em modelos ViT, destacando-se a técnica de *attention rollout* (Abnar; Zuidema, 2020). Essa metodologia foi originalmente proposta para aplicação em modelos tradicionais de *transformers*, permitindo quantificar o fluxo de atenção ao longo dos blocos de *transformers*.

Para cada bloco de *transformers* presente em um modelo, é possível obter uma matriz A_{ij} que representa a atenção que um *token* j direciona a um *token* i na camada seguinte. Esse fluxo é calculado por meio da multiplicação das matrizes A correspondentes a cada par de camadas dentro do bloco.

É importante destacar que os blocos de *transformers* também possuem conexões residuais que influenciam o fluxo da atenção. Portanto, para considerar a presença dessas conexões, soma-se a matriz identidade I à matriz de atenção A_{ij} , conforme indicado abaixo.

$$A'_{ij} = A_{ij} + I. \quad (8)$$

A matriz A'_{ij} representa o fluxo de atenção entre os *tokens* j e i de uma única *attention head*. Contudo, modelos ViT geralmente possuem múltiplas *attention heads* em cada camada. Dessa

forma, é possível combinar as diferentes matrizes de atenção por meio do cálculo da média, do mínimo ou do máximo elemento a elemento das matrizes A'_{ij} , obtendo assim uma matriz única AR_{ij} que representa a explicação final.

É importante destacar que a escolha da operação de fusão influencia diretamente a forma como as explicações são apresentadas, conforme ilustrado na Figura 10.

Figura 10 – Aplicação do método de *attention rollout* considerando a fusão das matrizes A'_{ij} a partir do cálculo da média e do mínimo.



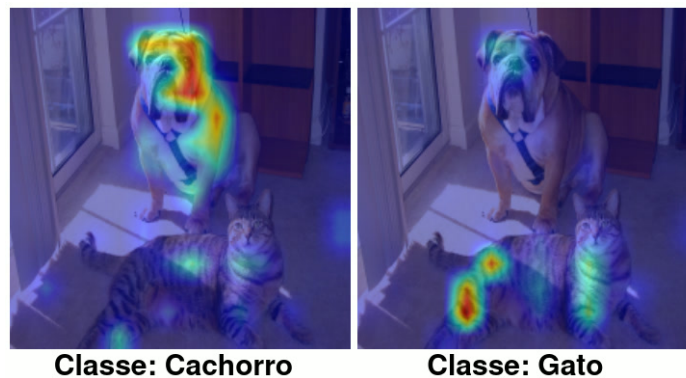
Fonte: Retirado de Gildenblat (2023).

Ao aplicar essa estratégia em modelos ViT, é comum observar a presença de ruído nas explicações, caracterizado pela ocorrência de múltiplas regiões da imagem destacadas como relevantes para a classificação final. Esse problema pode ser mitigado por meio da aplicação de diferentes métodos de fusão das matrizes A'_{ij} , além de um processo de descarte dos *pixels* com baixos valores de atenção.

Por fim, é importante ressaltar que este método não gera explicações específicas para uma classe determinada. Contudo, durante o cálculo das matrizes A'_{ij} , é possível incorporar informações de gradiente relativas a uma classe c , obtendo assim uma explicação AR_{ij}^c , conforme representado pela Equação 9. Na Figura 11 é ilustrado o resultado dessa operação.

$$AR_{ij}^c = A'_{ij} * grad_{ij}^c. \quad (9)$$

Figura 11 – Aplicação do método de *attention rollout* referente as classes de cachorro e gato.



Fonte: Adaptado de Gildenblat (2023).

2.6 MÉTRICAS PARA A AVALIAÇÃO DAS EXPLICAÇÕES

Uma das formas tradicionais de avaliar a capacidade explicativa de modelos de redes neurais é por meio de métricas qualitativas, que consistem em comparações visuais "lado a lado" entre propostas de aprimoramento e abordagens previamente consolidadas na literatura (Singh et al., 2021). No entanto, esse tipo de análise não permite conclusões objetivas, pois a interpretação dos resultados torna-se subjetiva. Por isso, torna-se fundamental definir um conjunto de métricas quantitativas capazes de avaliar com maior precisão as explicações geradas pelos modelos.

Nesse contexto, o estudo proposto por Poppi et al. (Poppi et al., 2021) apresenta um conjunto de métricas quantitativas para avaliação de explicações, além de oferecer uma pontuação única que facilita a comparação entre diferentes abordagens. Entre as métricas propostas destacam-se: coerência, complexidade, queda de confiança e *average DCC*.

2.6.1 Coerência (CO)

A métrica de Coerência (CO) avalia a estabilidade estrutural e a consistência interna das explicações visuais. Fundamentalmente, um mapa de ativação $CAM_c(x)$, associado a uma imagem x e a uma classe c , define-se como coerente se permanecer invariante quando utilizado para mascarar os dados de entrada. Em outras palavras, a explicação gerada para a imagem mascarada $x \odot CAM_c(x)$ deve aproximar-se da explicação original, satisfazendo a condição $CAM_c(x \odot CAM_c(x)) \approx CAM_c(x)$. Essa propriedade é formalmente expressa na Equação 10.

$$CAM_c(x \odot CAM_c(x)) \text{ igual a } CAM_c(x). \quad (10)$$

Para mensurar essa propriedade, calcula-se o coeficiente de correlação de Pearson entre os mapas de ativação original e transformado, conforme apresentado na Equação 11. O resultado é normalizado para o intervalo $[0, 1]$, no qual valores mais próximos de 1 indicam maior coerência e robustez da explicação (Riche et al., 2013; Cornia et al., 2016; Cornia et al., 2017).

$$CO(x) = \frac{Cov(CAM_c(x \odot CAM_c(x)), CAM_c(x))}{\sigma CAM_c(x \odot CAM_c(x)) \sigma CAM_c(x)}. \quad (11)$$

2.6.2 Complexidade (COM)

Já a métrica de complexidade (COM) quantifica a dispersão espacial do mapa de ativação. Mapas de alta complexidade tendem a apresentar ativação em regiões amplas e difusas, o que pode prejudicar a interpretabilidade clínica ao introduzir ambiguidade. Em contrapartida, explicações de baixa complexidade, que se concentram em áreas compactas e relevantes para o diagnóstico, são geralmente mais desejáveis. Para estimar esse comportamento, utiliza-se a norma L_1 do mapa de ativação, conforme formalizado na Equação 12.

$$COM(x) = \|CAM_c(x)\|_1. \quad (12)$$

Sendo assim, quanto menor o número de *pixels* com maior intensidade presentes em uma explicação, menor será sua complexidade, sendo o valor dessa métrica limitado ao intervalo $[0, 1]$.

2.6.3 Queda de Confiança (QC)

A métrica de queda de confiança (QC) mensura o quanto a confiança de predição do modelo diminui quando este é restringido apenas às regiões destacadas pela explicação. Sejam y_c a pontuação de predição na imagem completa e o_c na entrada mascarada. A métrica é definida como:

$$QC(x) = \max(0, y_c - o_c) / y_c. \quad (13)$$

Considerando que os valores estão representados no intervalo $[0, 1]$, valores mais baixos indicam que a explicação captura as regiões efetivamente responsáveis pela predição, preservando a confiança sob condições de entrada restritas. Dessa forma, uma menor QC implica maior fidelidade da explicação ao processo decisório interno do modelo (Soomro; Niaz; Choi, 2024).

2.6.4 Average DCC (ADCC)

Para consolidar o desempenho ao longo das três dimensões supracitadas, a métrica ADCC calcula a média harmônica entre a CO, $1 - COM$ e $1 - QC$:

$$ADCC(x) = 3 \left(\frac{1}{CO(x)} + \frac{1}{1 - COM(x)} + \frac{1}{1 - QC(x)} \right)^{-1}. \quad (14)$$

Essa métrica penaliza deficiências em qualquer aspecto isolado, assegurando que apenas explicações equilibradas e informativas obtenham pontuações elevadas. Representadas por valores no intervalo $[0, 1]$, pontuações de ADCC mais altas indicam que as explicações são consistentes, concisas e fidedignas, características fundamentais para a aplicação confiável em cenários clínicos e científicos.

2.7 MÉTRICAS PARA A AVALIAÇÃO DE DESEMPENHO DE MODELOS DE APRENDIZADO PROFUNDO

A avaliação rigorosa de modelos de aprendizado profundo aplicados ao diagnóstico médico exige uma análise que transcenda as métricas tradicionais de desempenho. No contexto da histopatologia digital, em que a confiabilidade clínica é tão crítica quanto a precisão diagnóstica, torna-se fundamental quantificar não apenas o que o modelo decide, mas também por que e como essa decisão é tomada (Ding et al., 2025; López et al., 2025; Albuquerque et al., 2024). Nesse cenário, estabelece-se um compromisso intrínseco entre desempenho e explicabilidade.

Historicamente, observa-se que o aumento na complexidade das arquiteturas, visando maior acurácia, tende a reduzir sua transparência, tornando-as "caixas-pretas" (Hassija et al., 2024). O objetivo desta avaliação é, portanto, verificar se os mecanismos de explicação propostos produzem mapas de atenção inteligíveis e robustos, bem como monitorar se a integração desses mecanismos preserva ou potencializa a capacidade preditiva do modelo. Idealmente, busca-se um equilíbrio em que a explicabilidade não comprometa o desempenho e funcione como indicativo de que o modelo aprendeu características morfologicamente relevantes (Wang; Popel; Sulam, 2024; Vásquez-Venegas et al., 2024).

Para quantificar a eficácia preditiva dos modelos, o processo de avaliação de desempenho supervisionado faz uso de métricas clássicas, como a acurácia (ACC), precisão (PR) e *F1-Score*. Essas métricas permitem mensurar a capacidade de discriminação entre classes e a robustez das predições, fornecendo uma visão que contempla não apenas a taxa global de acertos, mas também o equilíbrio entre a recuperação de instâncias relevantes e a confiabilidade das classificações positivas (Tharwat, 2020).

2.7.1 Acurácia (ACC)

A acurácia (ACC) avalia o desempenho global do modelo de classificação, quantificando a proporção de previsões corretas em relação ao universo total de amostras submetidas à inferência. Fundamentalmente, essa métrica indica a capacidade do classificador em atribuir o rótulo correto a uma imagem x , independentemente da classe a que ela pertença, refletindo a eficácia geral do aprendizado dos padrões visuais (Grandini; Bagli; Visani, 2020). A definição conceitual é dada pela razão entre o número de acertos e o total de instâncias avaliadas, conforme expresso na Equação 15.

$$ACC = \frac{1}{N} \sum_{i=1}^N 1(\hat{y}_i = y_i). \quad (15)$$

Para mensurar essa propriedade de forma analítica com base na matriz de confusão, consideram-se os Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). A formalização é apresentada na Equação 16. O resultado é definido no intervalo $[0, 1]$, no qual valores mais próximos de 1 denotam um modelo com alta assertividade na discriminação das classes presentes no conjunto de dados (Hossin; M.N, 2015; Powers, 2020).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \quad (16)$$

2.7.2 Precisão (PR)

A métrica de precisão (PR) avalia a confiabilidade das classificações positivas realizadas pelo modelo, focando na minimização de erros do tipo I (falsos positivos). Fundamentalmente, ela quantifica a proporção de instâncias corretamente identificadas como pertencentes a uma classe

de interesse em relação ao total de instâncias que o modelo predisse, correta ou incorretamente, como sendo dessa classe. Em cenários de diagnóstico médico, uma alta precisão é crucial para assegurar que o sistema não gere alarmes falsos excessivos, garantindo que uma predição positiva tenha alta probabilidade de corresponder à realidade clínica (Maier-Hein et al., 2024).

Para mensurar essa propriedade a partir da matriz de confusão, calcula-se a razão entre os VP e o total de predições positivas, composto pela soma de VP e FP , conforme apresentado na Equação 17. O resultado é definido no intervalo $[0, 1]$, no qual valores mais próximos de 1 indicam que o modelo possui alta especificidade preditiva, apresentando baixa incidência de falsas detecções (Powers, 2020; Sokolova; Lapalme, 2009).

$$PR = \frac{VP}{VP + FP}. \quad (17)$$

2.7.3 Recall (RC)

A métrica de *recall* (RC), frequentemente denominada Sensibilidade no domínio médico, avalia a capacidade do modelo em identificar corretamente todas as instâncias positivas presentes no conjunto de dados. Fundamentalmente, ela quantifica a taxa de acerto entre os casos que são verdadeiramente positivos (VP), focando na minimização de FN . No contexto de diagnóstico histopatológico, esta métrica assume importância crítica, pois a falha em detectar uma patologia existente (um FN) pode acarretar consequências severas ao prognóstico do paciente, atrasando intervenções terapêuticas necessárias (Maier-Hein et al., 2024).

Para mensurar essa propriedade a partir da matriz de confusão, calcula-se a razão entre os VP e o total de casos reais positivos, composto pela soma de VP e FN , conforme apresentado na Equação 18. O resultado é definido no intervalo $[0, 1]$, no qual valores mais próximos de 1 indicam que o modelo possui alta sensibilidade, sendo capaz de recuperar a vasta maioria das condições patológicas presentes nas amostras analisadas (Powers, 2020; Hossin; M.N, 2015).

$$RC = \frac{VP}{VP + FN}. \quad (18)$$

2.7.4 F1-score

O *F1-score* oferece uma avaliação equilibrada do desempenho do modelo ao calcular a média harmônica entre a PR e a RC . Essa métrica é essencial em cenários com desequilíbrios entre as classes ou situações em que os custos de FP e FN são igualmente críticos (Grandini; Bagli; Visani, 2020). Diferentemente da média aritmética, a média harmônica penaliza valores extremos, exigindo que o modelo apresente um bom desempenho simultâneo tanto na recuperação de instâncias relevantes quanto na minimização de falsos alarmes.

Para mensurar essa propriedade em termos dos componentes da matriz de confusão, utiliza-se a relação entre os VP , FP e FN , conforme apresentado na Equação 19. O resultado é definido no intervalo $[0, 1]$, no qual valores mais próximos de 1 indicam um classificador robusto, capaz

de manter um compromisso ideal entre precisão e sensibilidade (Powers, 2020; Hossin; M.N, 2015).

$$F1-Score = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}. \quad (19)$$

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentadas abordagens da literatura que exploram o uso de redes neurais convolucionais para a geração de explicações, abrangendo diferentes técnicas e contextos de aplicação. Além disso, são analisados trabalhos mais recentes que empregam modelos baseados em *vision transformers* e arquiteturas híbridas.

3.1 REDES NEURAIAS CONVOLUCIONAIS EXPLICÁVEIS

No campo da inteligência artificial explicável, um dos principais desafios consiste em encontrar um equilíbrio entre explicabilidade e acurácia, especialmente no contexto de modelos baseados em CNNs. Isso se deve ao fato de que, quanto mais explicável é uma rede, maior tende a ser o impacto negativo sobre sua acurácia (Bell et al., 2022). Diante disso, diferentes arquiteturas de CNNs foram propostas com o objetivo de alcançar um balanço adequado entre esses dois aspectos. Entre essas propostas, destaca-se o modelo *Explainable Convolutional Neural Network* (XCNN) (Tavanaei, 2020).

A arquitetura XCNN foi inspirada nos modelos de redes adversárias generativas, do inglês *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014), nos quais um componente gerador é responsável por criar novas imagens a partir dos dados de treinamento. De forma análoga, o modelo XCNN propõe a utilização de dois componentes principais voltados ao aprimoramento das explicações: um módulo *encoder-decoder* e uma CNN discriminadora.

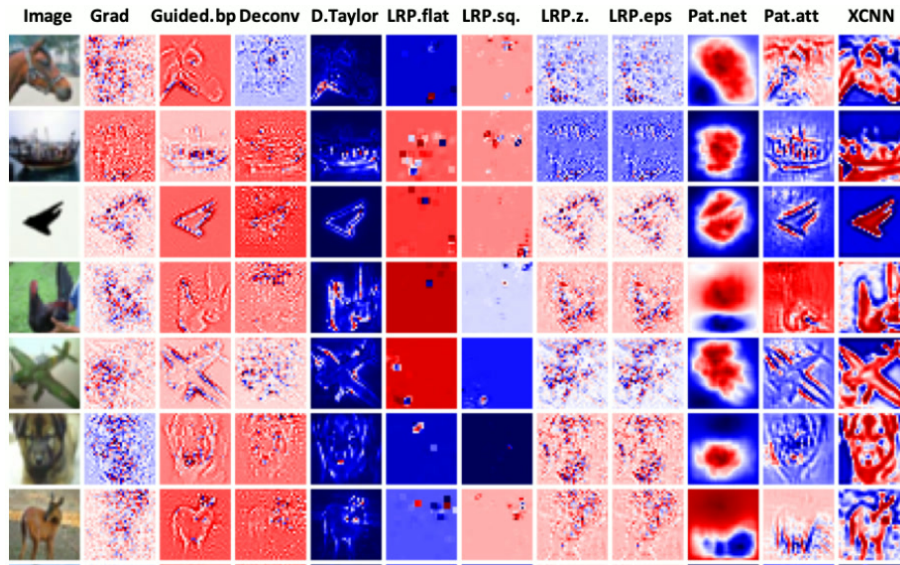
O módulo *encoder-decoder* tem a função de extrair características a partir de uma imagem de entrada, de modo que essas características sejam armazenadas em um vetor de características C . Esse vetor C é então utilizado como entrada em um processo de *upsampling*, cuja função é reconstruir a imagem original na forma de um mapa de calor, destacando as regiões mais relevantes da imagem. Esse mapa de calor é posteriormente fornecido como entrada para uma rede discriminadora (VGG-16), com o objetivo de realizar a classificação final da imagem.

Os resultados obtidos por essa arquitetura conseguiram superar diferentes técnicas voltadas à geração de explicações, tais como: *Gradient* (Simonyan; Vedaldi; Zisserman, 2014), *Guided Backpropagation*, *Deconvolution* (Springenberg et al., 2015), *Deep Taylor Decomposition* (Montavon et al., 2017), *Layer-wise Relevance Propagation* (LRP) (Bach et al., 2015), *PatternNet* e *PatternAttribution* (Kindermans et al., 2017). Na Figura 12 é apresentado os recortes dos resultados obtidos pelo modelo XCNN em comparação com essas diferentes abordagens de explicação.

É importante destacar que a comparação dos resultados obtidos pelo modelo XCNN em relação aos demais métodos presentes na literatura foi realizada por meio de métricas qualitativas, uma abordagem comumente adotada no contexto de XAI (Johs; Agosto; Weber, 2020).

Já o modelo proposto por Fukui et al. (Fukui et al., 2019) apresentou uma abordagem distinta

Figura 12 – Explicações obtidas pelo modelo XCNN em comparação aos métodos *gradiente*, *guided backprop*, *deconvolution*, *deep Taylor decomposition*, LRP, *PatternNet* e *PatternAttribute*.



Fonte: Adaptado de Tavanaei (2020).

para o aprimoramento das explicações geradas por CNNs. O modelo *Attention Branch Network* (ABN) utiliza arquiteturas CNN já consolidadas, como ResNet (He et al., 2016) e VGGNet (Simonyan; Zisserman, 2015a), como *backbones* para compor uma arquitetura organizada em três blocos principais: o *feature extractor*, o *attention branch* e o *perception branch*.

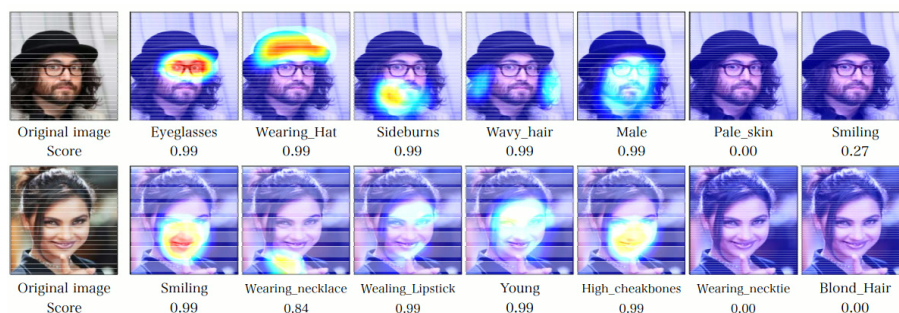
O *feature extractor* é composto por todos os blocos convolucionais (ou residuais) do modelo *backbone*, sendo responsável por extrair as características da imagem de entrada x_i na forma de mapas de características $g(x_i)$. Esses mapas são, então, fornecidos ao *attention branch*, módulo composto pelo último bloco convolucional (ou residual) do modelo *backbone*, que, além de gerar um rótulo $L_{att}(x_i)$ para os dados de entrada, também é responsável por construir um mapa de atenção (*attention map*) $M(x_i)$, o qual destaca as regiões mais relevantes da imagem.

O mapa de atenção é combinado com as características $g(x_i)$ por meio de um mecanismo de atenção, cujo resultado é utilizado como entrada para o *perception branch*, cuja estrutura de camadas é análoga à do *attention branch*. A função do *perception branch* é obter um segundo rótulo $L_{per}(x_i)$, de modo que os valores de *loss* associados aos rótulos $L_{att}(x_i)$ e $L_{per}(x_i)$ possam ser combinados em um único valor. Esse valor combinado é, então, utilizado para a atualização dos pesos do modelo durante o processo de treinamento.

Essa metodologia não apenas tornou o modelo mais explicável, como ilustrado na Figura 13, mas também proporcionou um desempenho superior em termos de acurácia quando comparado a modelos do estado da arte. No *dataset* CelebA (Liu et al., 2015), o modelo alcançou uma acurácia de 91,07%, superando os resultados obtidos por outras abordagens previamente propostas na literatura.

Quanto ao uso de métodos de XAI na área médica, o trabalho desenvolvido por Sarp

Figura 13 – Explicações obtidas pelo modelo ABN no *dataset* CelebA.



Fonte: Retirado de Fukui et al. (2019).

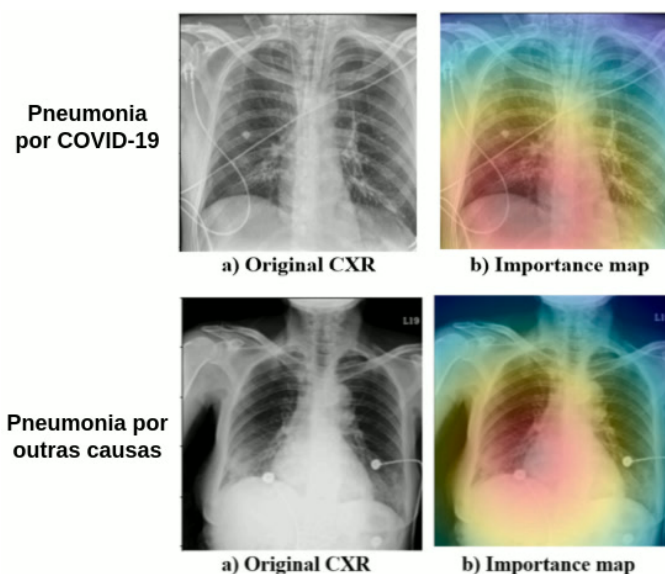
et al. (Sarp et al., 2023) propõe o uso de dois módulos convolucionais combinados com a técnica conhecida como *local interpretable model-agnostic explanations* (LIME) (Ribeiro; Singh; Guestrin, 2016) para a geração de explicações no contexto de imagens radiográficas pulmonares relacionadas à COVID-19. A metodologia proposta é composta por três etapas principais: segmentação do pulmão, classificação da presença de COVID-19 e geração de explicações.

A etapa de segmentação pulmonar é realizada por meio de um modelo convolucional híbrido, que utiliza a rede VGG11 (Simonyan; Zisserman, 2015a) como extrator de características, em conjunto com o modelo U-Net, responsável pela segmentação das regiões pulmonares nas imagens de raio-X. Em seguida, a etapa de classificação da imagem segmentada foi conduzida utilizando diferentes modelos de CNN, como ResNet e VGG, sendo que o modelo VGG16 apresentou o melhor desempenho, alcançando uma acurácia de 96% no conjunto de dados utilizado. Por fim, na etapa de geração de explicações, a técnica LIME foi empregada para produzir mapas de calor que destacam as regiões mais relevantes da imagem em casos positivos de COVID-19.

É importante destacar que, neste estudo, a avaliação da capacidade explicativa dos modelos foi realizada exclusivamente por meio de métricas qualitativas. Na Figura 14 são apresentadas duas amostras de explicações geradas, considerando casos de pneumonia causados por COVID-19 e por outras causas, respectivamente.

Já o trabalho elaborado por Iglesias et al. (Iglesias; Menendez; Talavera, 2025) propõe o MOC VAE, uma arquitetura baseada em *variational autoencoders* que explora as propriedades de completude e continuidade do espaço latente para gerar explicações de baixa intrusão. A metodologia apresenta três etapas principais: extração de descritores visuais e semânticos por meio do MOC VAE, que combina perda de reconstrução com um termo de classificação para adicionar *feedback* semântico ao espaço latente (Kingma; Welling, 2022); navegação adversarial nesse espaço por meio de um algoritmo genético SPEA2 (Deb et al., 2002), que otimiza simultaneamente a maximização da probabilidade da classe alvo, a minimização da quantidade de mutações e a minimização da magnitude das alterações; e reconstrução dos descritores perturbados para produzir mapas de saliência capazes de destacar as regiões mais relevantes das

Figura 14 – Explicações obtidas pelo modelo proposto por Sarp et al., considerando um caso de pneumonia causado pela COVID-19 e por outras causas, respectivamente.



Fonte: Adaptado de Sarp et al. (2023).

radiografias. Em contraste com abordagens baseadas em superpixels ou delimitação manual de regiões, esta proposta dispensa configuração de hiperparâmetros adicionais e assegura coerência semântica ao operar em um espaço de baixa dimensão.

Para avaliação, o modelo foi treinado e testado no conjunto Padchest, balanceado em quatro classes (normal, escoliose, alongação da aorta e cardiomegalia), com pré-processamento que incluiu segmentação pulmonar via TransResUNet (Reza; Amin; Hashem, 2020) e normalização das imagens para dimensão de 224×224 pixels. Na tarefa de classificação, alcançou acurácia de 61% e *F1-score* de 0,61, superando arquiteturas de referência como EfficientNetB0 (Tan; Le, 2019), DenseNet201 (Huang et al., 2017), VGG19 (Simonyan; Zisserman, 2015b) e ResNet50 (He et al., 2016). No ataque adversarial, obteve mudança média de confiança de 0,6382 para a classe-alvo e taxa de sucesso de 86,45% na fronteira de Pareto dos indivíduos gerados, conforme os objetivos de pesquisa RQ1 e RQ4. A análise da quantidade mínima de variáveis latentes a serem alteradas revelou que, em média, apenas 40,81 delas precisam ser modificadas, com magnitude total média de 6,45, para induzir erro de classificação, atendendo à questão RQ2.

Em relação à capacidade explicativa, as reconstruções resultantes destacaram corretamente as regiões anatômicas associadas às patologias em 80% dos casos (RQ3), superando métodos como LIME (Ribeiro; Singh; Guestrin, 2016), DeepCover (Golshanrad; Faghih, 2024), SAG e EGAE em até 56,39 pontos percentuais na precisão das áreas evidenciadas (RQ5). Assim, o MOC VAE estabelece um *framework* de XAI para radiografias que oferece explicações de baixo impacto e alinhadas semanticamente ao diagnóstico, sem necessidade de ajustes específicos de hiperparâmetros para o processo de explicação.

3.2 EXPLICABILIDADE EM MODELOS ViT

Modelos baseados em ViT introduziram avanços significativos no processo de geração de explicações (Kashefi et al., 2023). Isso se deve à capacidade desses modelos de analisarem os dados de entrada em um contexto global (Dosovitskiy et al., 2021), em contraste com as CNNs, que processam a imagem por meio da análise de pequenas regiões locais. Dessa forma, os modelos ViT são capazes de interpretar uma imagem em sua totalidade, o que os torna particularmente adequados para o contexto de análise de imagens médicas. Nessa área, alterações sutis em uma imagem podem representar diferenças clínicas relevantes, sendo capazes de modificar completamente o diagnóstico ou o prognóstico de uma condição de saúde.

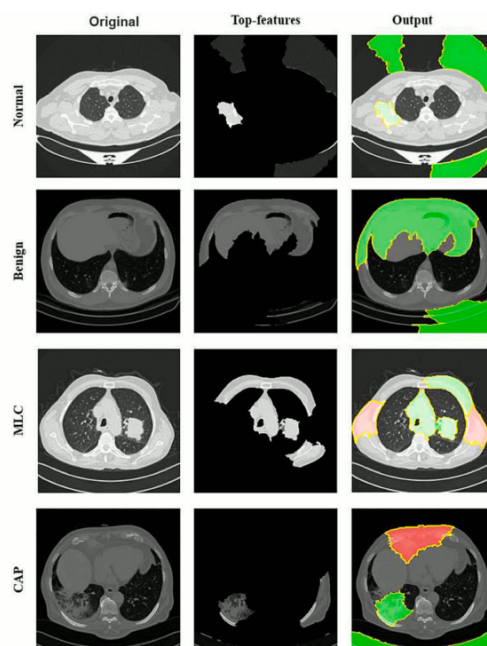
Neste contexto, o estudo apresentado por Islam et al. (Islam et al., 2024) propõe um modelo híbrido denominado *DCNN-ViT-GRU*, o qual visa combinar as características de CNNs, *Gated Recurrent Units* (GRUs) e ViTs para realizar a detecção e classificação de anomalias pulmonares causadas por diferentes vetores, como tumores, COVID-19 e pneumonia. Além disso, o estudo incorpora técnicas XAI, como o LIME e o *Shapley Additive Explanations* (SHAP), a fim de permitir uma melhor interpretação do comportamento do modelo (Ribeiro; Singh; Guestrin, 2016; Lundberg; Lee, 2017).

De forma geral, o modelo DCNN-ViT-GRU é constituído por três módulos principais: DCNN, ViT e GRU. O módulo DCNN apresenta uma estrutura similar à dos blocos convolucionais amplamente utilizados na literatura (Huang et al., 2017; He et al., 2016), nos quais camadas convolucionais são intercaladas com camadas de *pooling*, seguidas pela aplicação da função de ativação Tanh. Já o módulo ViT adota a arquitetura proposta por Dosovitskiy et al. (2021), incorporando componentes como *patch embeddings*, codificação posicional, blocos de atenção (*attention blocks*) e camadas totalmente conectadas (*fully connected layers*). Essa configuração aprimora a capacidade do modelo de processar imagens e extrair representações relevantes dos dados de entrada.

O terceiro módulo, baseado em unidades GRU, recebe como entrada os vetores de características extraídos pelos módulos DCNN e ViT. Esses vetores são combinados por meio de uma operação de soma, resultando na fusão das informações extraídas por ambas as arquiteturas. A sequência resultante é então processada por uma camada GRU, cuja função é capturar as dependências temporais entre as representações, permitindo ao modelo aprender relações sequenciais presentes nos dados. Por fim, o vetor gerado é submetido à uma função *softmax*, que produz as probabilidades associadas a cada classe.

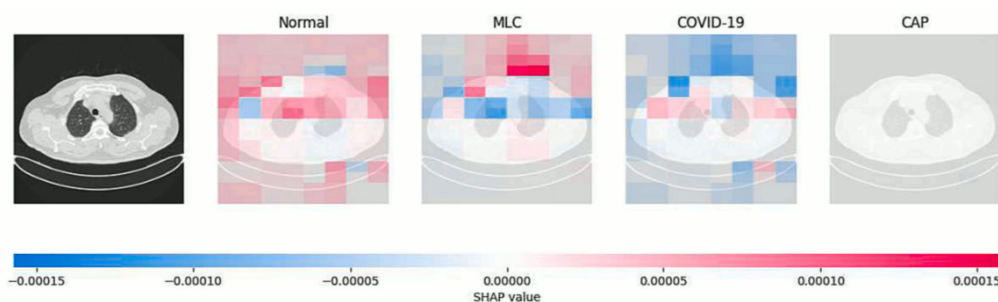
Em termos de desempenho, a proposta apresentada por Islam et al. (2024) alcançou acurácias de 99% e 99,86% em dois conjuntos de dados compostos por radiografias pulmonares. Além disso, o modelo apresentou resultados satisfatórios quanto à aplicação dos métodos explicáveis LIME e SHAP, conforme ilustrado respectivamente na Figura 15 e Figura 16. Cabe destacar que, neste estudo, a capacidade explicativa do modelo foi avaliada exclusivamente por meio de métricas qualitativas.

Figura 15 – Explicações obtidas pelo método LIME referentes às classes Normal, COVID-19, *Malignant Lung Cancer* e CAP.



Fonte: Retirado de Dosovitskiy et al. (2021).

Figura 16 – Explicações obtidas pelo método SHAP referentes às classes Normal, COVID-19, *Malignant Lung Cancer* e CAP.



Fonte: Retirado de Dosovitskiy et al. (2021).

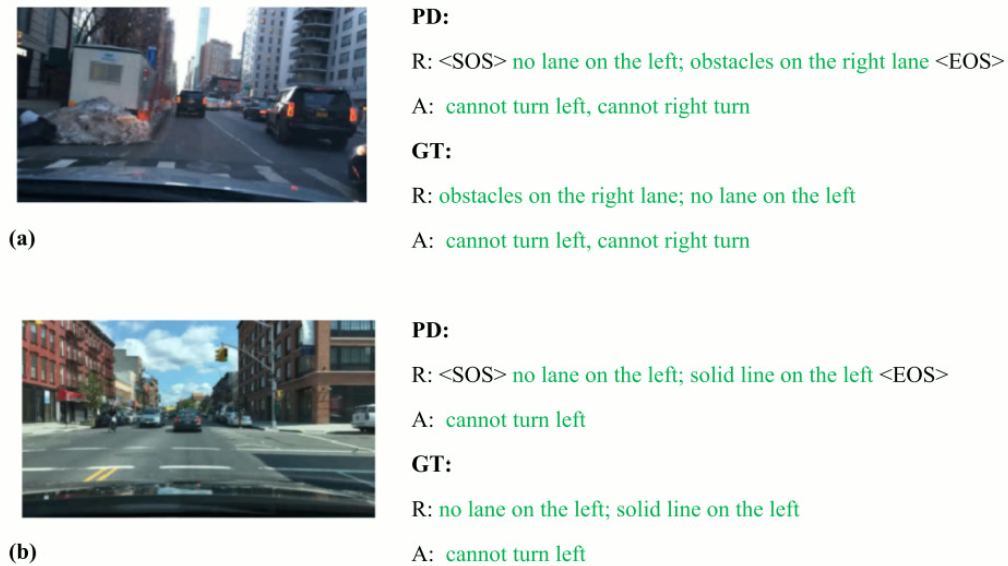
Já o trabalho desenvolvido por Dong et al. (2023) propõe o uso de modelos ViT e técnicas de XAI no contexto de desenvolvimento de sistemas para veículos autônomos. Nesse estudo, os autores abordam a confiabilidade na utilização de carros autônomos, destacando a importância de mecanismos explicáveis para a tomada de decisão. Assim, o sistema proposto foi capaz de fornecer explicações textuais e visuais que indicam como cada decisão foi tomada.

A metodologia apresentada emprega dois modelos distintos para a geração de explicações e a definição das ações do sistema. O primeiro modelo consiste em uma rede de *captioning*, responsável por gerar uma legenda descritiva a partir de uma imagem de entrada. Essa rede é composta por camadas convolucionais, utilizadas para extrair características vetoriais da imagem, as quais são então processadas por um modelo baseado em *transformers* para produzir uma descrição textual completa da cena.

A descrição gerada pela rede de *captioning* é, posteriormente, combinada com a imagem original e utilizada como entrada para um segundo modelo, baseado em ViT. Esse modelo realiza a fusão dos dados visuais e textuais, gerando uma representação final que descreve a ação executada, bem como a justificativa para tal decisão. A explicação é fornecida por meio de *attention maps*, os quais indicam visualmente as regiões da imagem que influenciaram a decisão do sistema.

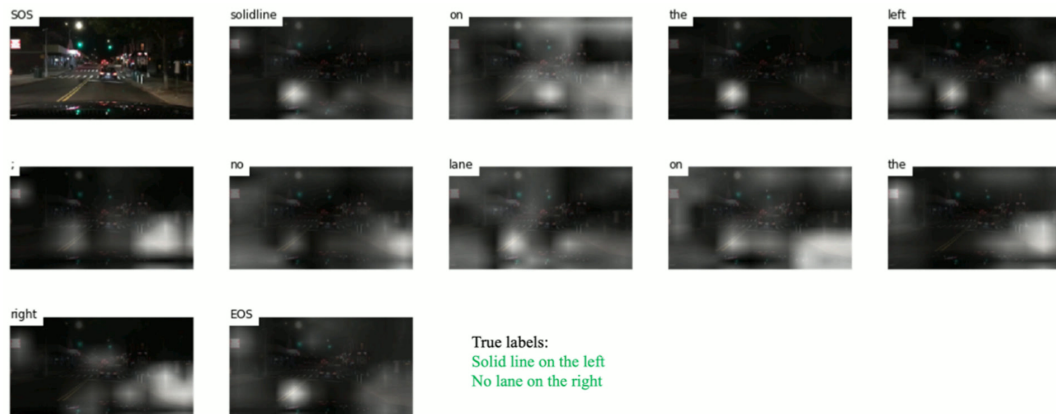
Entre os resultados obtidos, o estudo realizou uma série de análises qualitativas para avaliar a eficácia das explicações geradas. Na Figura 17 é apresentada a descrição textual da ação realizada e os motivos associados, enquanto que na Figura 18 são ilustrados os *attention maps* correspondentes às regiões consideradas mais relevantes pela rede para sua classificação final.

Figura 17 – Descrição da explicação referente a tomada de decisão, em que, PD é a previsão, GT é o *ground truth*, R é a razão e A é a ação.



Fonte: Retirado de Dong et al. (2023).

Figura 18 – Exemplos dos mapas de atenção gerados pela metodologia proposta Dong et al. (2023).



Fonte: Retirado de Dong et al. (2023).

3.3 CONSIDERAÇÕES SOBRE OS TRABALHOS CORRELATOS

As abordagens discutidas neste capítulo evidenciam a diversidade de estratégias empregadas para incorporar explicabilidade em modelos de redes neurais, com destaque para arquiteturas baseadas em convoluções e em ViT. No caso das redes convolucionais, observa-se um esforço contínuo para equilibrar desempenho e interpretabilidade, como demonstrado pelos modelos XCNN (Tavanaei, 2020) e ABN (Fukui et al., 2019). Essas propostas adotam mecanismos específicos como módulos *encoder-decoder* ou ramificações baseadas em mecanismos de atenção que permitem destacar visualmente regiões relevantes nas imagens de entrada, oferecendo aos usuários uma compreensão mais clara sobre o processo decisório da rede.

Por outro lado, os modelos baseados em ViT têm se destacado por sua capacidade de manipular relações espaciais complexas por meio de mecanismos de atenção. Estudos recentes, como os de Islam et al. (2024) e Dong et al. (2023), ilustram como essa arquitetura pode ser combinada a métodos XAI como LIME, SHAP e *attention maps* para fornecer explicações visuais e textuais robustas. Em particular, a proposta de Dong et al. incorpora um modelo de *captioning* para enriquecer semanticamente as explicações, ampliando a transparência em sistemas de tomada de decisão autônomos.

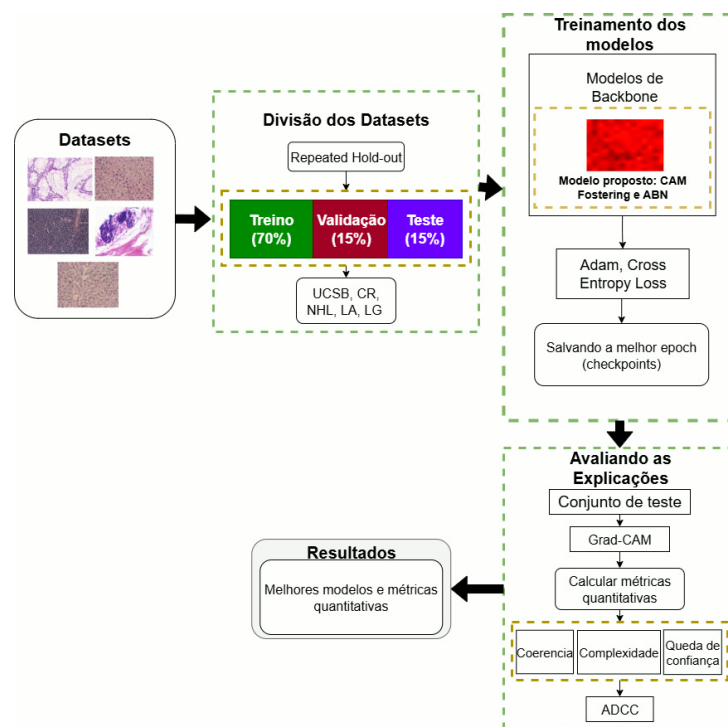
Apesar dos avanços, nota-se que a maioria dos trabalhos ainda se apoia fortemente em métricas qualitativas para avaliar a qualidade das explicações geradas. Essa limitação compromete a objetividade na comparação entre métodos e dificulta a padronização na avaliação da interpretabilidade. Assim, a necessidade de métricas quantitativas mais robustas permanece como um desafio central na área de inteligência artificial explicável, sobretudo quando aplicada a domínios sensíveis, como o setor médico ou o controle de veículos autônomos.

Nesse contexto, observa-se uma tendência promissora na combinação de diferentes paradigmas arquiteturais, como DCNNs, ViTs e redes recorrentes, com técnicas de XAI, visando tanto o aumento da acurácia quanto a transparência dos modelos. Essa convergência representa um caminho relevante para futuras investigações, sobretudo no desenvolvimento de sistemas críticos em que a explicação das decisões é tão essencial quanto a sua precisão.

4 METODOLOGIA

A metodologia desenvolvida neste trabalho foi organizada em três etapas principais. Na primeira, os cinco *datasets* de imagens histológicas foram particionados por meio da estratégia de *repeated hold-out* (Tanner; Bornehag; Gennings, 2019), utilizando uma divisão de 70/15/15 ao longo de cinco iterações, de modo a garantir maior robustez na avaliação. Na segunda etapa, diferentes modelos convolucionais e híbridos foram treinados e analisados tanto em suas versões *baseline* quanto em *backbones* do modelo proposto. Cada arquitetura foi treinada com os *datasets* previamente particionados, e a *F1-score* foi empregada para selecionar o melhor ponto de treinamento entre as *epochs*. Por fim, na terceira etapa, os modelos treinados foram utilizados para gerar mapas de explicação por meio da técnica Grad-CAM, os quais foram avaliados por um conjunto de métricas quantitativas com o objetivo de mensurar o poder explicativo de cada arquitetura. Na Figura 19 é apresentada uma visão geral da metodologia proposta.

Figura 19 – Diagrama esquemático do método proposto.



Fonte: Elaborado pelo autor.

4.1 DEFINIÇÃO E PARTICIONAMENTO DOS DATASETS H&E

Este trabalho fez uso de cinco *datasets* que representam quatro tipos de tecidos histológicos, sendo que todos os *datasets* possuem amostras pigmentadas com hematoxilina e eosina. O primeiro *dataset* (UCSB) considera imagens de câncer de mama, fornecidas pela Universidade

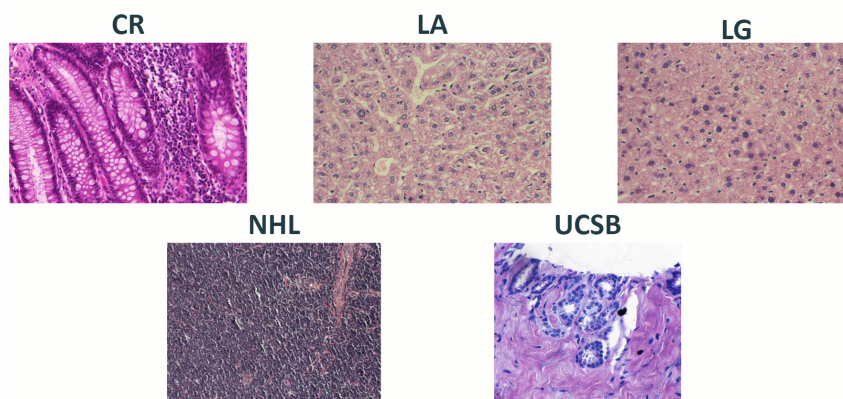
da Califórnia em Santa Bárbara (Gelasca et al., 2008). Este *dataset* é composto por 58 amostras divididas em duas classes: benigno (38) e maligno (26).

O segundo *dataset* (CR) é composto por imagens de tecido colorretal (Sirinukunwattana et al., 2017), totalizando 165 amostras divididas entre duas classes: benigno (74) e maligno (91). Essas imagens foram obtidas via áreas histológicas fotografadas utilizando um *Zeiss MIRAX MIDI Slide Scanner*, em que foi utilizada uma escala de resolução de $0.620 \mu m$, o que é equivalente a uma ampliação de 20x.

O terceiro *dataset* (NHL) foi publicado pelo *National Cancer Institute* em colaboração com o *National Institute on Aging* (Shamir et al., 2008). Este conjunto de imagens é composto por 173 amostras de linfomas não Hodgkin divididas em três classes: MCL — *mantle cell lymphoma* (99); FL — *follicular lymphoma* (62); e CLL — *chronic lymphocyte leukemia* (12). As imagens que compõem este conjunto foram capturadas por um microscópio *Zeiss Axioscope* com uma ampliação de 20x e com uma câmera digital *Axion Cam MR5*. As imagens obtidas por esse processo foram armazenadas sem compressão, com uma resolução de $1388 \times 1040 \text{ pixels}$, taxa de quantização de 24 bits e modelo de cores RGB.

Por fim, os quarto e quinto conjuntos de amostras foram fornecidos pela *Atlas of Gene Expression in Mouse Ageing Project* (AGEMAP) (Zahn et al., 2007). As imagens foram adquiridas a partir de um microscópio *Carl Zeiss Axiovert 200* com uma ampliação de 40x. O quarto *dataset* (LG) é composto por 265 amostras de tecidos hepáticos coletados de ratos submetidos a uma dieta restrita em calorias, sendo que dos animais analisados: 150 machos e 115 fêmeas; o quinto *dataset* (LA) é composto por 529 imagens divididas em quatro classes, em que cada classe representa um grupo de idade distinto de camundongos submetidos a uma dieta *ad libitum*, sendo as classes: um mês (100); seis meses (115); 16 meses (162) e 24 meses (152). Na Figura 20 são ilustradas algumas amostras presentes em cada *dataset*, enquanto na Tabela 1 apresenta-se uma visão geral sobre todos os *datasets*.

Figura 20 – Exemplos de imagens histológicas presentes em cada *dataset*.



Fonte: Elaborado pelo autor.

Para garantir a consistência dos resultados em relação ao poder explicativo de cada modelo, a estratégia de *repeated hold-out* foi aplicada em cada *dataset* individualmente. Para este processo,

Tabela 1 – Visão geral de todos os cinco *datasets*.

Dataset	Tipo de tecido	Classes	Amostras	Resolução
UCSB	Tumores de mama	2	58	896×768
CR	Tumores colorretais	2	165	Entre 567×430 e 775×522
NHL	Linfomas não-Hodgkin	3	173	Entre 86×65 e 1388×1040
LG	Tecido hepático	2	265	417×312
LA	Tecido hepático	4	529	417×312

Fonte: Elaborado pelo autor.

foi aplicado um particionamento 70/15/15 em cada conjunto de dados ao longo de cinco iterações, em que: 70% do *dataset* foi dedicado ao processo de treino; 15% para o processo de validação e 15% para teste. Vale ressaltar que as imagens para compor cada subconjunto foram selecionadas aleatoriamente de cada conjunto original.

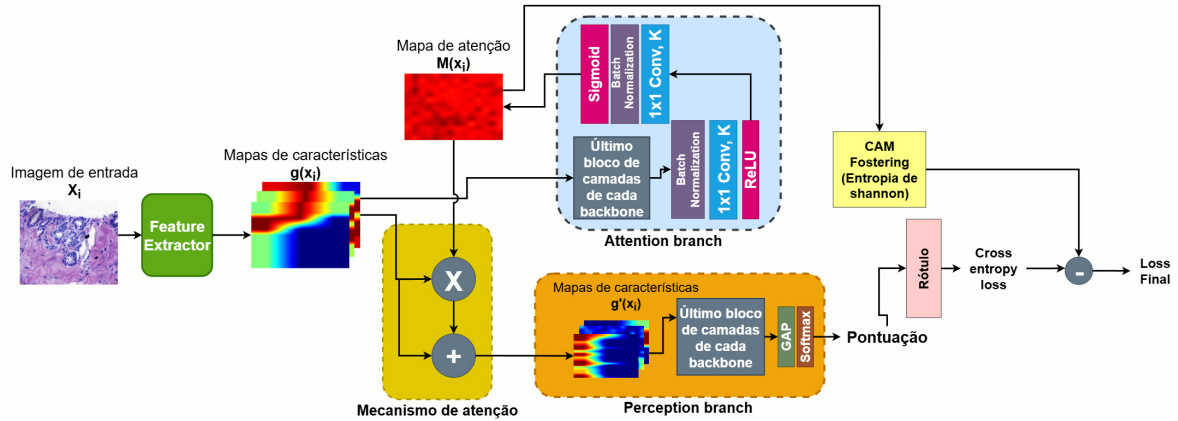
4.2 DEFINIÇÃO E TREINAMENTO DO MODELO PROPOSTO: ABN COM *CAM FOSTERING*

O modelo desenvolvido neste trabalho foi construído a partir da arquitetura ABN (Fukui et al., 2019), integrada à estratégia *CAM Fostering* (Schöttl, 2022), visando aprimorar a capacidade explicativa de redes neurais convolucionais e híbridas. O esquema computacional contemplou diferentes arquiteturas de referência, incluindo ResNet-50 (He et al., 2016), DenseNet-201 (Huang et al., 2017), EfficientNet-b0 (Tan; Le, 2019), ResNeXt-50 (Xie et al., 2017) e ConvNeXt (Woo et al., 2023), além dos modelos híbridos CoatNet-small (Dai et al., 2021) e Uniformer (Li et al., 2023). A proposta final foi organizada em três módulos, *feature extractor*, *attention branch* e *perception branch*, e incorporou dois mecanismos voltados ao fortalecimento das explicações geradas: o *attention mechanism* e o *CAM fostering*. É importante ressaltar que, para todos os modelos *backbones* utilizados nesta metodologia, o módulo *attention branch* foi composto exclusivamente pelo último bloco de camadas convolucionais de cada modelo. Na Figura 21 é apresentado um diagrama esquemático do modelo proposto.

4.2.1 *Feature Extractor*

O módulo *feature extractor* é composto por todas as camadas (convolucionais ou ViT) de cada modelo *backbone* utilizado (ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNeXt e CoatNet-small), excluindo os estágios finais dedicados à classificação. Sua função é transformar a imagem de entrada X_i em um conjunto de mapas de características $g(X_i)$ capazes de capturar padrões espaciais hierárquicos inerentes às imagens histológicas, tais como granularidade de textura, morfologia celular e arquitetura tecidual (Tavanaei, 2020). Esses mapas de características representam uma codificação rica e semanticamente densa, fundamental para suportar tanto a interpretabilidade quanto as tarefas de classificação do modelo.

Figura 21 – Diagrama esquemático do modelo proposto utilizando camadas convolucionais para a composição do *attention branch*.



Fonte: Elaborado pelo autor.

Os mapas de características extraídos por este módulo são encaminhados simultaneamente para os outros dois módulos principais da arquitetura: o *attention branch* e o *perception branch*. O *attention branch* é projetado para gerar mapas de atenção espacial $M(X_i)$ que destacam regiões relevantes para a classe. Esses mapas são utilizados pelo mecanismo de atenção para orientar o refinamento das características e, concomitantemente, pela estratégia *CAM Fostering* para regularizar a distribuição da atenção por meio de restrições baseadas em entropia. Esse uso duplo promove a consistência entre as áreas da imagem que impulsionam as previsões do modelo e aquelas apresentadas como explicações. Uma descrição detalhada do ramo de atenção é fornecida na próxima subseção.

Paralelamente, o *perception branch* aplica um mecanismo de atenção que modula os mapas de características originais utilizando os mapas de atenção, produzindo uma representação refinada $g'(X_i)$ focada em áreas relevantes para o diagnóstico. Esse design incentiva um alinhamento funcional entre a explicação e a tomada de decisão, mitigando limitações conhecidas em abordagens de XAI anteriores em imagens médicas, as quais frequentemente tratam a interpretabilidade como um processo *post hoc* ou externo (Ioannidis; Maniadis, 2023; Xu et al., 2019).

4.2.2 Attention Branch

O módulo *attention branch* recebe como entrada os mapas de características $g(X_i)$ gerados pelo *feature extractor* e produz um mapa de atenção $M(X_i)$. Posteriormente, esse mapa é utilizado para modular as características de classificação e calcular o termo de entropia na estratégia *CAM Fostering*, aprimorando a interpretabilidade do modelo durante o treinamento (Misra et al., 2021). Estruturalmente, o *attention branch* é constituído pelo último bloco convolucional de cada *backbone*, o qual produz um tensor com dimensões $K \times w \times h$, onde K representa o número de canais e $w \times h$ a resolução espacial. Esse tensor é submetido a uma sequência composta por *batch normalization*, uma camada convolucional 1×1 e uma

função de ativação ReLU, reduzindo a dimensionalidade para um mapa intermediário de canal único. Uma segunda sequência de normalização e ativação, composta por outra camada de *batch normalization*, uma convolução 1×1 e uma função sigmoide, é aplicada para gerar o mapa de atenção final $M(X_i)$, restringindo seus valores ao intervalo $[0, 1]$.

É importante ressaltar que, diferente do modelo ABN original, o módulo de *attention branch* proposto não possui um bloco de classificação para o mapa de atenção. Essa modificação foi realizada para permitir a implementação da estratégia de *CAM fostering*.

4.2.3 Mecanismo de Atenção

O mecanismo de atenção implementado no *perception branch* segue a formulação estabelecida no modelo ABN original (Fukui et al., 2019), no qual o mapa de atenção $M(X_i)$ é utilizado para modular os mapas de características originais $g(X_i)$, produzindo um conjunto refinado de características $g'(X_i)$. Esse processo enfatiza regiões consideradas relevantes para a previsão, aprimorando a interpretabilidade e a eficácia da classificação final. O mecanismo é definido pela Equação 1:

$$g'(X_i) = (g(X_i) \times M(X_i)) + g(X_i). \quad (1)$$

Essa equação combina atenção elemento a elemento com aprendizado residual. O termo de atenção realça seletivamente as regiões salientes, enquanto a conexão residual preserva o contexto original completo das características, promovendo estabilidade representacional e melhorando o fluxo de gradiente durante o treinamento.

4.2.4 Perception Branch

O módulo *perception branch* foi responsável por gerar a saída de classificação final do modelo. Este módulo recebe como entrada os mapas de características aprimorados $g'(X_i)$ produzidos pelo mecanismo de atenção, integrando tanto a representação original das características quanto a orientação espacial proveniente dos mapas de atenção. Este módulo reutiliza o último bloco de cada arquitetura *backbone* original, preservando a abstração semântica inerente ao seu design.

Após esse bloco, uma camada de GAP é aplicada para comprimir as dimensões espaciais de cada mapa de características em um único valor escalar. Essa operação transforma o tensor $K \times w \times h$ em um vetor de dimensão K , no qual cada valor representa a ativação global de um canal correspondente. Essa representação é subsequentemente passada por uma função de ativação *softmax* para gerar as probabilidades de classe normalizadas.

A adoção da camada de GAP ao invés de uma *fully-connected* como utilizada no modelo ABN original, atende a um duplo propósito. Primeiro, preserva a correspondência espacial das características convolucionais, crucial para manter a interpretabilidade, ao evitar a perda de informações espaciais (Zhou et al., 2016). Segundo, reduz o número de parâmetros treináveis, minimizando assim o risco de *overfitting*. Essa escolha arquitetural assegura que as regiões

discriminativas identificadas pelo mecanismo de atenção permaneçam diretamente ligadas ao resultado final da classificação, reforçando a capacidade do modelo de produzir explicações espacialmente coerentes e clinicamente relevantes (Zhou et al., 2015).

4.2.5 CAM Fostering

Para aprimorar ainda mais a interpretabilidade do modelo proposto, a estratégia de *CAM Fostering* (Schöttl, 2022) foi integrada como um mecanismo auxiliar durante o treinamento. Essa abordagem introduz uma restrição baseada na teoria da informação sobre os mapas de atenção, incentivando a geração de padrões de ativação que não sejam nem excessivamente esparsos nem demasiadamente difusos.

O mecanismo opera por meio do cálculo da entropia de Shannon ce do mapa de atenção $M(X_i)$, a qual quantifica a diversidade das ativações no domínio espacial. Mapas de atenção com ativações altamente uniformes apresentam baixa entropia, indicando capacidade de localização limitada, ao passo que mapas com respostas espaciais diversificadas exibem maior entropia, sugerindo um conteúdo explicativo mais rico. A entropia ce é formalmente definida na Equação 2:

$$ce(M(X_i)) = - \sum M(X_i)_{ij} \ln M(X_i)_{ij}. \quad (2)$$

Os índices ij percorrem o domínio espacial bidimensional do mapa de atenção. Durante o treinamento, esse valor de entropia é incorporado à *loss function* como um termo de regularização, ponderado por um fator $\gamma_e \in [0, 10]$. Conforme sugerido na formulação original (Schöttl, 2022), valores mais elevados de γ_e amplificam a influência da regularização por entropia, aprimorando a qualidade da explicação, ainda que ao custo potencial de uma redução na acurácia da classificação. O objetivo final de treinamento l'_n é definido como a perda de entropia cruzada original l_n subtraída do termo de regularização ponderado pela entropia, conforme apresentado na Equação 3:

$$l'_n = l_n - \gamma_e * ce(M(X_i)). \quad (3)$$

Neste trabalho, a técnica *CAM Fostering* foi aplicada aos mapas de atenção $M(X_i)$ gerados pelo *attention branch* de cada modelo. Utiliza-se um fator de regularização $\gamma_e = 10$ para maximizar o efeito regularizador, assegurando a geração de mapas de atenção espacialmente mais diversos e informativos. A função de *cross-entropy loss* (Mao; Mohri; Zhong, 2023) permanece como o critério de otimização primário, ao passo que o termo *CAM Fostering* é empregado como uma restrição complementar para equilibrar o desempenho de classificação com a qualidade da explicação.

4.2.6 Treinamento dos Modelos

Para essa etapa, foram treinados os modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNeXt, CoatNet-small e Uniformer em suas versões *baseline* e aplicados como *backbones* para o modelo proposto. Esse processo foi importante para viabilizar uma comparação do impacto do modelo proposto quanto à qualidade das explicações em comparação aos modelos *baseline*.

Para acelerar o processo de treinamento, além de evitar problemas como *overfitting* dado o número reduzido de amostras em alguns *datasets*, a estratégia de *transfer learning* foi aplicada (Zhuang et al., 2019), sendo assim, todos os modelos estavam pré-treinados no conjunto de imagens ImageNet (Russakovsky et al., 2015), possibilitando um processo de ajuste fino dos pesos com um número reduzido de *epochs*.

O treinamento foi realizado em um total de 20 *epochs*, considerando uma taxa de aprendizado de 0,0001 e um *batch size* de 16. É importante ressaltar que, visando evitar *overfitting*, também foi aplicada a estratégia de *checkpoints* de forma que, a partir do cálculo da métrica *F1-score* durante a etapa de validação de cada *epoch*, foi selecionado apenas o treinamento que obteve a melhor métrica *F1-score* ao longo das 20 *epochs*. Para a atualização dos pesos foi utilizado o otimizador Adam, essa escolha se deve a capacidade desse otimizador de alcançar uma rápida convergência com um número reduzido de *epochs* (Kingma; Ba, 2014). A *loss function* escolhida para o treinamento foi a *cross-entropy* (Mao; Mohri; Zhong, 2023), que também foi utilizada em conjunto com a estratégia de *CAM fostering* para o treinamento dos modelos propostos por essa metodologia.

4.3 DEFINIÇÃO DO PROCESSO DE AVALIAÇÃO DAS EXPLICAÇÕES

Nessa etapa, foi aplicada uma série de métricas quantitativas que permitiram a avaliação dos mapas de ativação obtidos pelo método *Grad-CAM* dos modelos treinados na etapa 2. Para isso, o conjunto de teste de cada *dataset* obtido na etapa 1 foi utilizado para a criação dos mapas. As métricas que foram utilizadas para avaliar a capacidade explicativa de cada modelo estão descritas em detalhes nas Seções 2.6.1 até 2.6.4, que incluem: coerência (CO); complexidade (COM); queda de confiança (QC) e média DCC (ADCC).

A métrica CO propõe que um mapa de ativação obtido por uma rede neural não deve ser alterado se o mesmo mapa for fornecido como entrada para essa rede. O grau de diferença desses mapas é então medido em um intervalo $[0,1]$, em que quanto mais próximo de um, maior é a semelhança entre os mapas obtidos (Seção 2.6.1).

Já a métrica COM indica a quantidade de informações presentes em um mapa de ativação em que, considerando o intervalo $[0,1]$, quanto mais próximo é de um, maior é a complexidade do mapa; logo, ele apresenta uma capacidade explicativa menos relevante (Seção 2.6.2).

A métrica QC avalia o impacto na confiança de classificação de uma rede neural em que é fornecida como entrada a imagem referente à explicação gerada pela rede. Essa métrica é

limitada no intervalo $[0,1]$, em que, quanto mais próxima de um, maior é o impacto na confiança do modelo (Seção 2.6.3).

Por fim, a métrica ADCC foi responsável por fornecer uma única pontuação capaz de indicar o poder explicativo de um determinado modelo. Essa pontuação é calculada a partir da média harmônica entre as demais métricas, sendo assim, ela também é limitada ao intervalo $[0,1]$ em que, quanto mais próxima de um, melhor é o poder explicativo do modelo (Seção 2.6.4).

4.3.1 Pacotes de Software e Ambiente de Execução

A implementação da metodologia proposta foi realizada utilizando a linguagem de programação Python, versão 3.12.3. O desenvolvimento e o treinamento do modelo foram conduzidos utilizando a estrutura de aprendizado profundo PyTorch (Paszke et al., 2019), em combinação com a biblioteca PyTorch-Ignite (Fomin et al., 2020), que foi utilizada para otimizar as rotinas de treinamento e avaliação. As métricas de desempenho da classificação foram calculadas utilizando a biblioteca Scikit-learn (Pedregosa et al., 2011), enquanto todos os cálculos relacionados com métricas de explicação, tais como coerência, complexidade, queda de confiança e entropia, foram realizados utilizando NumPy (Harris et al., 2020). Além disso, todos os experimentos foram executados em um ambiente Windows, em uma máquina equipada com um processador Intel Core i9-14900HX, 32 GB de RAM e GPU NVIDIA RTX 5060 com 8 GB de memória dedicada.

5 RESULTADOS E DISCUSSÃO

Neste capítulo, apresenta-se uma avaliação sistemática da metodologia proposta por meio de um protocolo experimental em três etapas. Cada etapa foi projetada para avaliar aspectos distintos do desempenho e da interpretabilidade do modelo proposto por meio de métricas quantitativas e qualitativas.

Na primeira etapa (Seção 5.1), os modelos *backbone* originais, ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNeXt e CoatNet-small, foram avaliados usando as métricas de explicabilidade descritas na Seção 2.6: CO, COM, QC e ADCC. É importante ressaltar que as métricas apresentadas neste capítulo representam a média das mesmas obtidas ao longo de cinco interações por meio do processo de *repeated hold-out*. Esse método permitiu a realização de comparações entre os modelos *baselines* e os resultados apresentados pela metodologia proposta neste trabalho. Analogamente, na segunda etapa (Seção 5.2), a arquitetura proposta foi avaliada utilizando cada *backbone* individualmente. Os modelos modificados foram reavaliados usando as mesmas métricas da etapa anterior, possibilitando avaliar o impacto quantitativo das modificações propostas na qualidade das explicações geradas. Na última etapa (Seção 5.2.1), foi realizada uma análise visual comparativa das explicações geradas. Os mapas de calor Grad-CAM dos modelos originais e modificados foram justapostos para ilustrar qualitativamente as melhorias na interpretabilidade em amostras representativas dos conjuntos de dados histológicos.

5.1 MODELOS BACKBONE ORIGINAIS

Nas Tabelas 2 a 6 são apresentadas as métricas de explicabilidade, CO, COM, QC e ADCC para os modelos de *backbones* originais avaliados em cinco conjuntos de dados histológicos, apresentando as médias obtidas ao longo de cinco interações do processo de *repeated hold-out*. Nesses resultados, valores mais altos de CO e ADCC são preferíveis ($\uparrow$), enquanto valores mais baixos de COM e QC são desejáveis ($\downarrow$). É importante observar que todas as métricas são representadas em formato percentual para uma melhor interpretação dos resultados. A avaliação individual por *dataset* permite uma análise detalhada da interpretabilidade do modelo em diferentes domínios histológicos, destacando a influência das características arquitetônicas na qualidade das explicações baseadas em saliência. Esses resultados também servem como base para avaliar os ganhos de interpretabilidade alcançados posteriormente pela arquitetura proposta.

Entre os modelos avaliados, o CoatNet-small apresentou uma métrica ADCC consistentemente relevante em todos os *datasets*, alcançando desempenhos significativos nos *datasets* LA (72,75%), NHL (70,69%) e LG (65,39%). Essa tendência sugere um forte alinhamento entre as classes previstas e as regiões espaciais destacadas pelo Grad-CAM, indicando que a arquitetura híbrida do modelo, combinando camadas convolucionais com blocos ViT, permite explicações semanticamente mais coerentes e espacialmente significativas. Os campos receptivos globais das

Tabela 2 – Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* UCSB, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	23,26	0,10	24,16	45,60
DenseNet-201	27,62	0,10	10,15	52,10
EfficientNet-b0	27,95	0,11	22,57	53,82
ResNext-50	24,68	0,11	18,60	47,31
ConvNext	27,88	0,11	10,29	53,56
CoatNet-small	30,87	0,11	10,43	57,34
Uniformer	22,60	0,11	32,38	47,62

Fonte: Elaborado pelo autor.

Tabela 3 – Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* NHL, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	29,84	0,07	47,83	58,77
DenseNet-201	23,68	0,07	62,94	46,77
EfficientNet-b0	26,21	0,07	63,59	53,59
ResNext-50	24,64	0,07	58,17	48,97
ConvNext	21,67	0,07	30,45	46,13
CoatNet-small	35,71	0,07	55,13	70,69
Uniformer	23,34	0,07	66,05	51,78

Fonte: Elaborado pelo autor.

Tabela 4 – Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* CR, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	29,80	0,13	10,82	56,07
DenseNet-201	30,54	0,13	8,32	56,62
EfficientNet-b0	26,98	0,13	25,74	53,08
ResNext-50	33,65	0,13	7,96	60,02
ConvNext	24,68	0,13	13,13	49,05
CoatNet-small	34,43	0,13	13,73	62,01
Uniformer	23,07	0,13	35,07	49,53

Fonte: Elaborado pelo autor.

Tabela 5 – Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* LG, incluindo coerência (CO), complexidade (COM), queda de confiança (QC), e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	31,60	0,24	31,37	60,52
DenseNet-201	30,75	0,24	32,00	57,86
EfficientNet-b0	31,29	0,24	42,79	61,41
ResNext-50	30,75	0,24	18,13	57,55
ConvNext	25,89	0,24	6,72	49,77
CoatNet-small	33,03	0,24	45,13	65,39
Uniformer	23,51	0,24	51,78	49,98

Fonte: Elaborado pelo autor.

Tabela 6 – Métricas de explicabilidade dos modelos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* LA, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	23,59	0,24	67,49	47,54
DenseNet-201	28,73	0,24	65,64	58,91
EfficientNet-b0	32,38	0,24	69,30	65,59
ResNext-50	24,96	0,24	71,52	52,14
ConvNext	24,95	0,24	37,30	50,83
CoatNet-small	35,54	0,24	72,05	72,75
Uniformer	26,77	0,24	78,91	59,80

Fonte: Elaborado pelo autor.

camadas ViT são particularmente benéficos nesses contextos histopatológicos, onde estruturas relevantes podem ser não contíguas e dispersas pela imagem (Vaswani et al., 2023).

Em contrapartida, o modelo Uniformer, embora também adote uma arquitetura híbrida, apresentou limitações significativas na geração de explicações fiéis. Além de exibir valores de ADCC substancialmente mais baixos que o CoatNet em todos os *datasets* (por exemplo, 51,78% no NHL e 49,98% no LG), o modelo mostrou um comportamento desfavorável nas demais métricas. Notadamente, o Uniformer registrou os piores índices de QC nas cinco bases, atingindo valores elevados como 78,91% no LA e 66,05% no NHL, além de apresentar níveis consistentemente reduzidos de CO. Um possível motivo para esse desempenho inferior, apesar de o Uniformer também empregar mecanismos de atenção, está relacionado à forma como sua arquitetura integra convoluções e *self-attention*. Diferentemente de abordagens como o CoatNet, que mantêm uma hierarquia clara entre estágios convolucionais iniciais e blocos atencionais posteriores, favorecendo simultaneamente a captura de padrões locais e a agregação de contexto global, o Uniformer utiliza um desenho mais homogêneo, no qual a transição entre operações espaciais e atencionais é menos definida. Em cenários histopatológicos, essa menor especialização pode restringir a capacidade do modelo de realçar regiões morfolologicamente distintas ou sutilmente separadas no espaço, reduzindo a precisão das ativações produzidas pelo

Grad-CAM. Além disso, a integração mais difusa entre atenção e convolução pode levar o modelo a depender excessivamente de padrões globais, dificultando a identificação de microestruturas relevantes para a coerência e para a estabilidade das explicações, o que resulta em métricas menos competitivas quando comparadas às de modelos convolucionais consolidados ou de híbridos com maior hierarquização interna.

No grupo dos modelos estritamente convolucionais, o EfficientNet-b0 se destacou, alcançando valores elevados de ADCC em bases como nos *datasets* LA (65,59%) e LG (61,41%). Apesar de ser o modelo mais compacto entre os avaliados, sua arquitetura eficiente favorece a construção de hierarquias de características estáveis e interpretáveis, indicando que a profundidade ou número de parâmetros não são, por si só, determinantes da qualidade explicativa. O ResNeXt-50 e o ConvNeXt apresentaram desempenhos moderados, mas consistentes, com o ResNeXt-50 obtendo valores competitivos em CR (60,02%) e LG (57,55%), o que reflete o impacto positivo de arquiteturas modulares e estratégias aprimoradas de agregação de características. Ainda assim, seus mapas ativados tendem a apresentar menor precisão espacial em comparação com modelos híbridos.

O DenseNet-201, por sua vez, mostrou comportamento heterogêneo. Embora tenha alcançado boa CO em alguns cenários, como UCSB (CO de 27,62%) e LA (CO de 28,73%), o modelo apresentou valores de ADCC inferiores em conjuntos como CR (56,62%) e NHL (46,77%). Esses resultados sugerem que, apesar da reutilização intensiva de características promovida pela conectividade densa, a ausência de mecanismos explícitos de atenção pode limitar a capacidade do modelo de gerar explicações espacialmente focadas e semanticamente alinhadas à predição.

De modo geral, os resultados apresentados nas Tabelas 2 a 6 evidenciam uma variação significativa na capacidade explicativa das arquiteturas analisadas. As diferenças observadas reforçam a importância de investigar soluções que incorporem mecanismos orientados à atenção, regularização sensível à entropia ou estratégias arquiteturais que promovam alinhamento espacial mais consistente entre as regiões salientadas e os padrões decisórios dos modelos. Tais aspectos são fundamentais para aplicações em domínios sensíveis como análise de tecidos histológicos, nas quais a interpretabilidade não é apenas desejável, mas essencial para apoio clínico confiável.

5.2 AVALIAÇÃO DOS MODELOS PROPOSTOS

Nas Tabelas 7 a 11 são indicados os valores percentuais das métricas de explicabilidade obtidos após a aplicação do modelo proposto, que combina supervisão de atenção e regularização baseada em entropia (*CAM Fostering*) sobre cada *backbone* em todos os conjuntos de dados. Para cada configuração, são reportados os valores CO, COM, QC e ADCC. Nesse contexto, valores mais altos de CO e ADCC indicam explicações mais consistentes e alinhadas às regiões relevantes ($\uparrow$), enquanto valores mais baixos de COM e QC refletem mapas de ativação mais concisos e estáveis ($\downarrow$). Assim como nos resultados obtidos para os modelos base, todas as métricas são expressas em porcentagem.

Tabela 7 – Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* UCSB, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	31,72	0,11	22,66	59,11
DenseNet-201	32,31	0,11	15,35	58,54
EfficientNet-b0	29,20	0,11	21,60	55,56
ResNext-50	32,79	0,11	27,04	61,90
ConvNext	33,29	0,10	9,60	59,83
CoatNet-small	32,36	0,11	20,00	60,83
Uniformer	25,60	0,11	25,47	51,29

Fonte: Elaborado pelo autor.

Tabela 8 – Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* NHL, incluindo incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	34,23	0,07	64,01	70,01
DenseNet-201	20,37	0,07	66,70	42,40
EfficientNet-b0	27,87	0,07	63,57	57,46
ResNext-50	25,50	0,07	64,98	54,93
ConvNext	33,46	0,07	5,96	59,98
CoatNet-small	34,29	0,07	50,14	67,90
Uniformer	24,98	0,07	58,04	53,92

Fonte: Elaborado pelo autor.

Tabela 9 – Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* CR, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	32,05	0,13	14,39	58,64
DenseNet-201	27,91	0,12	23,25	51,18
EfficientNet-b0	25,07	0,13	27,98	51,08
ResNext-50	32,88	0,13	12,99	59,88
ConvNext	32,21	0,13	2,22	57,62
CoatNet-small	34,15	0,13	11,76	61,62
Uniformer	25,17	0,13	49,29	52,98

Fonte: Elaborado pelo autor.

Tabela 10 – Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* LG, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	30,85	0,24	45,76	62,32
DenseNet-201	33,71	0,24	34,88	63,36
EfficientNet-b0	30,84	0,24	45,21	60,41
ResNext-50	29,18	0,24	48,0	58,63
ConvNext	30,48	0,24	3,86	55,69
CoatNet-small	36,00	0,24	47,86	69,39
Uniformer	26,17	0,24	37,45	54,85

Fonte: Elaborado pelo autor.

Tabela 11 – Métricas de explicabilidade dos modelos propostos ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNext, CoatNet-small e Uniformer em relação ao *dataset* LA, incluindo CO, COM, QC e ADCC.

Modelo	CO↑	COM↓	QC↓	ADCC↑
ResNet-50	31,91	0,24	77,79	65,99
DenseNet-201	29,63	0,24	63,29	58,37
EfficientNet-b0	30,22	0,24	69,42	62,58
ResNext-50	36,30	0,24	78,47	75,27
ConvNext	31,16	0,24	7,93	57,33
CoatNet-small	37,89	0,24	69,33	76,54
Uniformer	28,16	0,24	75,29	61,03

Fonte: Elaborado pelo autor.

No conjunto de dados da UCSB (Tabela 7), o ResNext-50 alcançou o maior ADCC (61,90%), enquanto o ConvNext apresentou o maior CO (33,29%) e a menor QC (9,60%). O CoatNet-small também apresentou uma elevada média de ADCC (60,83%), indicando que a estratégia proposta melhora a interpretabilidade mesmo em cenários de menor escala. Para o conjunto de dados NHL (Tabela 8), o ResNet-50 apresentou o maior ADCC (70,01%), acompanhado por uma CO de 34,23%, sendo seguido de perto pelo CoatNet-small, que alcançou média de ADCC de 67,90% e CO de 34,29%. Notavelmente, o ConvNext registrou a menor QC (5,96%), reforçando sua capacidade de gerar explicações estáveis em amostras visualmente complexas. No conjunto de dados CR (Tabela 9), o CoatNet-small alcançou o maior CO (34,15%) e ADCC (61,62%). Paralelamente, o ConvNext mais uma vez demonstrou a menor QC (2,22%), sugerindo que as arquiteturas se beneficiam da abordagem proposta em contextos visuais mais simples. No conjunto de dados LG (Tabela 10), o CoatNet-small novamente liderou em ADCC (69,39%) e CO (36,0%), enquanto o ConvNext apresentou a menor QC (3,86%), confirmando sua robustez em morfologias de tecidos heterogêneos. Para o conjunto de dados LA (Tabela 11), o CoatNet-small alcançou o ADCC mais alto (76,54%) e o CO (37,89%). O ConvNext manteve a QC mais baixa (7,93), destacando sua consistência na produção de mapas de atenção estáveis em conjuntos de dados em grande escala e ricos em padrões. Em relação à generalização e implantação, as

arquiteturas propostas demonstraram robustez em conjuntos de dados com complexidade e escala variadas. O CoatNet-small obteve consistentemente os maiores ganhos de ADCC e Coerência, especialmente em conjuntos de dados morfológicamente complexos como NHL, LG e LA. Por outro lado, o ConvNext se destacou universalmente por manter a métrica QC minimamente baixa em todos os cinco *datasets*, indicando uma capacidade superior de gerar explicações estáveis e confiáveis. Em cenários de menor escala (UCSB), múltiplos modelos como ResNext-50 e ConvNext mostraram melhorias significativas, comprovando a adequação da solução proposta. Nesse contexto, os ganhos consistentes em interpretabilidade, particularmente o desempenho superior do CoatNet-small em ADCC e a notável estabilidade (baixa QC) do ConvNext em todos os cenários, validam a eficácia da abordagem proposta. Os resultados sugerem que a metodologia aplicada melhora a explicabilidade independentemente da arquitetura do modelo ou da complexidade do conjunto de dados.

5.2.1 Análise Qualitativa das Explicações

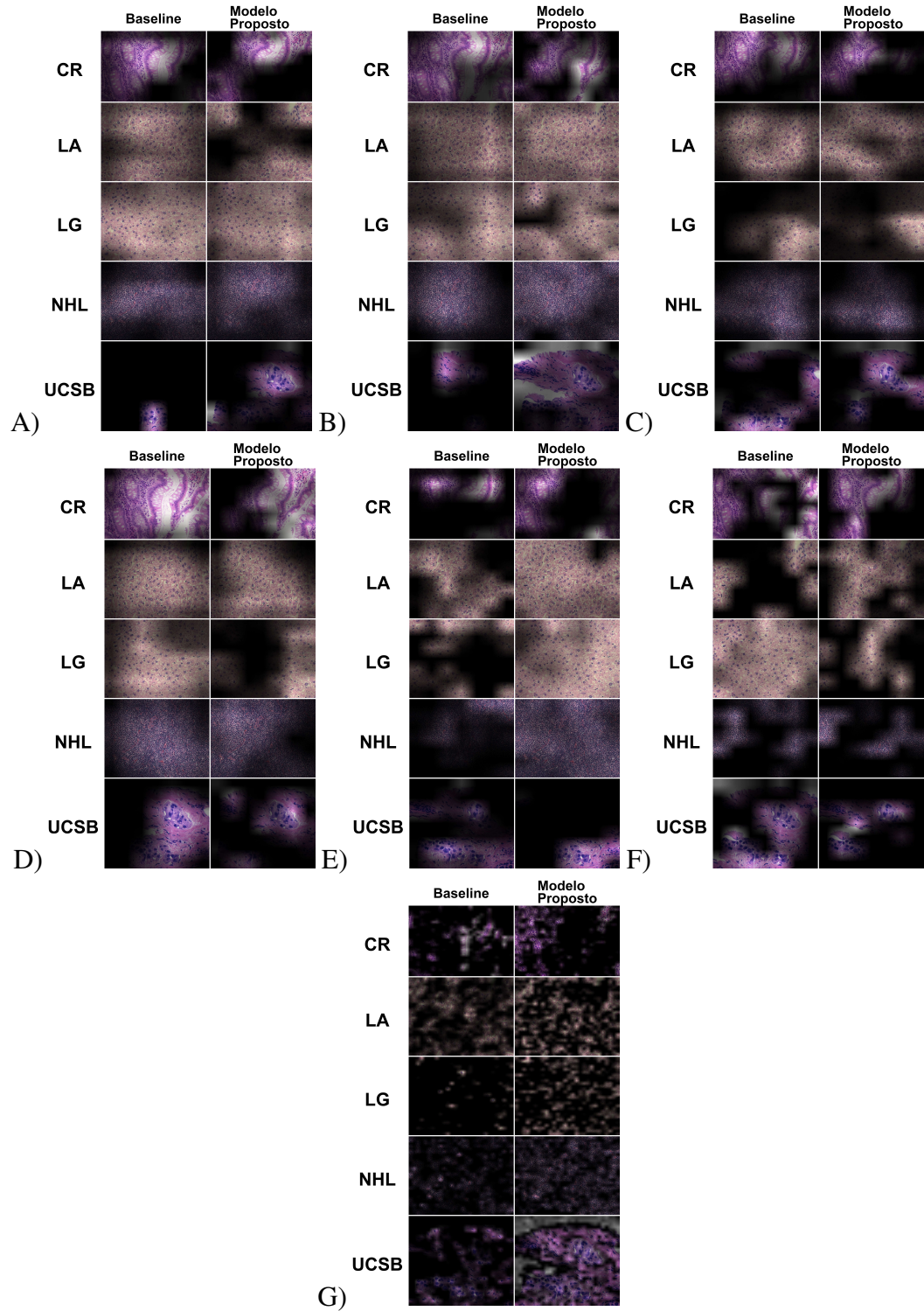
Na Figura 22 é fornecida uma comparação qualitativa das explicações Grad-CAM geradas pelos modelos originais e propostos para cada *backbone*. Para cada arquitetura, é selecionada uma imagem representativa, permitindo a inspeção visual das melhorias de interpretabilidade alcançadas por meio da supervisão da atenção e do *CAM Fostering*.

Na maioria das estruturas, os modelos propostos produzem explicações mais concentradas espacialmente, alinhadas semanticamente e relevantes para o diagnóstico. Essa melhoria é particularmente evidente no ResNet-50 (A) e no ResNeXt-50 (D), onde os modelos *baseline* exibem atenção difusa e inconsistente em grandes áreas da imagem. Com a arquitetura proposta, os mapas de atenção ficam focados em regiões de tecido relevantes para a classe, melhorando a interpretabilidade. Ganhos quantitativos reforçam essas observações visuais; o ResNeXt-50, por exemplo, registrou um aumento de 17,96% no ADCC no conjunto de dados NHL, o que é consistente com as melhorias visuais, nas quais os mapas de atenção delineiam claramente a morfologia do tumor.

O CoatNet-small (F) apresentou uma melhoria visual expressiva. Em sua versão *baseline*, suas explicações eram frequentemente amplas e imprecisas, cobrindo regiões irrelevantes. Com a metodologia proposta, as ativações passaram a se concentrar em padrões nucleares histologicamente relevantes, especialmente nos conjuntos NHL e LA. Esse refinamento é coerente com seu ADCC superior (77,90% no NHL) e com a maior média geral de ADCC (67,71%), reforçando o impacto positivo da integração arquitetônica adotada. Além disso, o aumento na coerência (40,05% no NHL) indica explicações mais consistentes e menos sensíveis a artefatos visuais.

O EfficientNet-b0 (C), por outro lado, apresentou um comportamento mais discreto. Embora seu desempenho *baseline* fosse robusto em conjuntos de menor complexidade, como o UCSB, as melhorias propostas não resultaram em ganhos visuais substanciais. Em alguns casos, como no NHL, o mapa de calor modificado permaneceu visualmente semelhante ao *baseline*. Os resultados quantitativos refletem esse cenário, com ADCC praticamente estável (de 57,50% para

Figura 22 – Comparação visual das explicações geradas pelo método Grad-CAM produzidas pelos modelos *baseline* (esquerda) e proposto (direita), em que: (A) ResNet-50, (B) DenseNet-201, (C) EfficientNet-b0, (D) ResNeXt-50, (E) ConvNeXt (F) CoatNet-small e (G) Uniformer.



Fonte: Elaborado pelo autor.

57,42%), sugerindo que sua arquitetura compacta pode limitar os benefícios da regularização adicional.

O Uniformer (G) apresentou o contraste mais acentuado. Sua versão *baseline* exibiu explicações altamente difusas e ruidosas, sem qualquer localização útil na maioria dos conjuntos (especialmente LA, LG e NHL). A metodologia proposta não elevou o Uniformer ao desempenho das melhores arquiteturas, mas mitigou de forma clara essas deficiências. Os mapas de ativação se tornaram menos ruidosos e passaram a apresentar concentrações mais discerníveis, com maior destaque no UCSB, onde o foco tornou-se mais nítido e com menos artefatos. Esses achados são compatíveis com os resultados quantitativos, que apontam redução da métrica QC (de 52,84% para 49,11%) e aumento do ADCC (de 51,74% para 54,81%). Esses resultados indicam que a regularização pode atenuar limitações de interpretabilidade mesmo em arquiteturas cuja atenção apresenta integração menos hierárquica.

A partir dessas análises visuais e quantitativas, é possível notar que o modelo proposto melhora sistematicamente a interpretabilidade, sobretudo em arquiteturas mais profundas ou híbridas. Ao alinhar a atenção espacial com uma supervisão sensível à entropia, os mapas de ativação tornam-se mais localizados, discriminativos e clinicamente significativos. Isso posiciona o modelo como uma ferramenta promissora para a interpretação de imagens histológicas, promovendo tanto maior transparência quanto maior rigor explicativo, sem comprometer a capacidade preditiva.

5.2.2 Resumo Comparativo entre os Modelos Propostos e *Baseline*

Nas Tabelas 12 e 13 são apresentados resumos comparativos a partir das métricas médias de explicabilidade em todos os conjuntos de dados, considerando cada *backbone* em sua forma original e após a aplicação da estratégia proposta.

Tabela 12 – Comparação das médias das métricas de explicabilidade (CO e COM) antes e depois da aplicação da metodologia proposta.

Modelo	CO↑		COM↓	
	<i>Baseline</i>	Proposto	<i>Baseline</i>	Proposto
ResNet-50	27,62	32,15	0,16	0,16
DenseNet-201	28,26	28,79	0,16	0,16
EfficientNet-b0	28,96	28,64	0,16	0,16
ResNeXt-50	27,73	31,33	0,16	0,16
ConvNeXt	25,02	32,12	0,16	0,16
CoatNet-small	33,91	34,94	0,16	0,16
Uniformer	23,86	26,02	0,16	0,16

Fonte: Elaborado pelo autor.

Considerando esta visão geral comparativa (Tabela 12 e Tabela 13), o ResNet-50 apresentou a melhoria mais substancial, com o ADCC médio aumentando de 53,70% para 63,21%, representando um ganho de 9,51%. De forma similar, o ResNeXt-50 também se beneficiou

Tabela 13 – Comparação das médias das métricas de explicabilidade (QC e ADCC) antes e depois da aplicação da metodologia proposta.

Modelo	QC↓		ADCC↑	
	Baseline	Proposto	Baseline	Proposto
ResNet-50	36,33	44,92	53,70	63,21
DenseNet-201	35,81	40,70	54,45	54,77
EfficientNet-b0	44,80	45,56	57,50	57,42
ResNeXt-50	34,87	46,30	53,20	62,12
ConvNeXt	19,58	5,91	49,86	58,09
CoatNet-small	39,29	39,82	65,64	67,25
Uniformer	52,84	49,11	51,74	54,81

Fonte: Elaborado pelo autor.

robustamente, com o ADCC crescendo 8,92% (de 53,20% para 62,12%). Ambos os modelos também indicaram aumentos significativos na métrica CO, destacando a vantagem de incorporar a metodologia proposta em arquiteturas convolucionais que não possuem modelagem de contexto global integrada.

O ConvNeXt demonstrou um ganho de ADCC de 8,23% (de 49,86% para 58,09%), mas sua melhoria mais notável foi na métrica QC, que foi drasticamente reduzida de 19,58% para 5,91% (uma melhoria de 13,67%), validando a capacidade do método em gerar explicações mais estáveis.

É também particularmente notável o impacto sobre o Uniformer. Conforme observado anteriormente, o modelo *baseline* apresentava limitações severas, registrando os piores índices de QC (média de 52,84%) e baixa Coerência (23,86%). A metodologia proposta reverteu parcialmente essa tendência: o ADCC médio aumentou de 51,74 para 54,81 (+3,07%), o CO subiu para 26,02 (+2,16%) e, crucialmente, a instabilidade (QC) foi reduzida para 49,11% (uma melhoria de 3,73%). Embora partindo de uma base baixa, essas melhorias indicam que a regularização foi capaz de mitigar as fraquezas de fidelidade e estabilidade inerentes à arquitetura original do Uniformer.

Em contrapartida, o CoatNet-small, já sendo a *baseline* mais forte (ADCC 65,64%), obteve um ganho mais contido de 1,61% (para 67,25%). Os modelos EfficientNet-b0 (de 57,5 para 57,42) e DenseNet-201 (de 54,45 para 54,77) mantiveram-se relativamente estáveis, sugerindo que seus designs altamente otimizados podem restringir a influência da regularização adicional.

No geral, a estratégia proposta melhora consistentemente a interpretabilidade na maioria das arquiteturas, aumentando a coerência e a métrica combinada ADCC, independentemente do projeto arquitetônico ou da escala do conjunto de dados. Os ganhos observados, desde a recuperação de modelos com limitações (Uniformer) até o aprimoramento de arquiteturas clássicas (ResNet-50), reforçam sua aplicabilidade em diversos domínios histológicos.

5.3 VISÃO GERAL SOBRE OS DESEMPENHOS DE CLASSIFICAÇÃO

Embora o objetivo principal deste trabalho seja melhorar a interpretabilidade dos modelos de aprendizado profundo por meio de modificações arquitetônicas, também é importante avaliar se essas mudanças afetam o desempenho da classificação. A Tabela 14 apresenta uma visão geral da pontuação média da *F1-score* e da Acurácia (ACC) (%) em todos os conjuntos de dados para cada *backbone*, destacando os resultados de classificação alcançados pelos modelos *baselines* em comparação com os da arquitetura proposta.

Tabela 14 – Média de ACC e *F1-score* (%) de todos os *datasets* para os modelos *baseline* e propostos.

Modelo	ACC		<i>F1-score</i>	
	<i>Baseline</i>	Proposto	<i>Baseline</i>	Proposto
ResNet-50	94,29	92,35	94,95	93,17
DenseNet-201	93,47	94,55	94,44	95,24
EfficientNet-b0	94,22	93,92	94,97	94,72
ResNeXt-50	93,60	91,59	94,22	92,61
ConvNeXt	95,31	95,55	95,50	96,13
CoatNet-small	91,24	94,50	91,31	95,14
Uniformer	95,61	93,44	96,06	94,37

Fonte: Elaborado pelo autor.

Uma análise mais detalhada sobre os dados contidos na Tabela 14 reforça a viabilidade e o sucesso da metodologia proposta, indicando que ganhos substanciais em interpretabilidade são alcançados sem comprometer significativamente o desempenho de classificação. Os ganhos mais expressivos foram observados no CoatNet-small, que registrou um aumento de 3,26% na ACC (de 91,24% para 94,5%) e 3,83% na *F1-score* (de 91,31% para 95,14%). Tendências positivas claras também foram vistas no DenseNet-201 (*F1-score* incremento de 0,80%) e no ConvNeXt (incremento de 0,63% na *F1-score*). Esses resultados sugerem que a regularização imposta pelos mecanismos de explicabilidade pode, na verdade, beneficiar a capacidade de generalização desses modelos. Nos demais modelos, as métricas permaneceram altamente estáveis, confirmando que a metodologia pode ser integrada sem custos de precisão. O EfficientNet-b0, por exemplo, apresentou uma variação de -0,30% na ACC e -0,25% na *F1-score*, o que na prática representa uma total preservação de seu desempenho. Mesmo nas arquiteturas que apresentaram leves flutuações negativas, como o ResNet-50 (-1,78% na *F1-score*), ResNeXt-50 (-1,61%) e Uniformer (-1,69%), as variações são marginais e o desempenho geral permanece em um nível competitivo. Em resumo, a análise de classificação valida a proposta. Em quatro dos sete modelos (CoatNet, DenseNet, ConvNext e EfficientNet), o desempenho preditivo foi marginalmente aprimorado ou preservado. Nos casos com leves ajustes, a troca é favorável, aceitando uma variação marginal (inferior a 2,20%) em troca dos ganhos substanciais em interpretabilidade, estabilidade e redução de ruído visual. Isso confirma a contribuição da arquitetura proposta para cenários clínicos, em que a explicabilidade é tão crítica quanto a precisão.

6 CONCLUSÃO

Este estudo apresentou uma metodologia que integra mecanismo de atenção com o regularizador baseado em entropia (*CAM Fostering*) para melhorar a explicabilidade de modelos de redes neurais no contexto de imagens histopatológicas. Por meio de experimentos abrangentes em sete modelos *backbone* (ResNet-50, DenseNet-201, EfficientNet-b0, ResNeXt-50, ConvNeXt, CoatNet-small e Uniformer) e cinco conjuntos de dados de imagens histológicas H&E, o método proposto alcançou ganhos consistentes na métrica de explicabilidade combinada (ADCC) para seis das sete arquiteturas. A análise comparativa apresentada nas Tabelas 12 e 13 revelou melhorias substanciais, como o aumento de 9,51% no ADCC médio do ResNet-50 e 8,92% no ResNeXt-50. Além disso, a estabilidade das explicações foi drasticamente aprimorada, notavelmente no ConvNeXt, que reduziu sua métrica QC média de 19,58% para 5,91%. Crucialmente, o desempenho da classificação foi preservado ou melhorado na maioria dos modelos, com o CoatNet-small registrando um aumento de 3,83% na *F1-score*, enquanto as demais arquiteturas (Tabela 14) apresentaram apenas variações marginais (inferiores a 2,20%), confirmando que os ganhos de interpretabilidade não exigem um sacrifício de precisão.

O *framework* proposto oferece três contribuições principais: um design modular compatível com *backbones* convolucionais e híbridos; uma *loss function* de treinamento ciente da entropia que afasta os mapas de atenção de padrões excessivamente estreitos ou difusos, gerando visualizações Grad-CAM mais claras e confiáveis; e um conjunto de avaliação quantitativa baseado em métricas de coerência, complexidade, queda de confiança e ADCC, permitindo a avaliação objetiva de mapas de saliência em diferentes modelos e conjuntos de dados.

Ao integrar a regularização baseada em entropia e supervisão da atenção espacial, nossa abordagem destaca consistentemente regiões relevantes para o diagnóstico, sem comprometer a precisão preditiva. A capacidade de aprimorar tanto arquiteturas clássicas (ResNet-50) quanto remediar *baselines* com limitações severas (Uniformer), mantendo a estabilidade preditiva, valida a metodologia. Este trabalho, portanto, oferece uma solução capaz de aumentar a transparência e a confiança no diagnóstico histopatológico assistido por modelos de aprendizado profundo.

Como proposta de trabalhos futuros, é possível aprimorar a arquitetura integrando módulos ViT diretamente à *attention branch*, com o objetivo de aproveitar sua capacidade de capturar dependências de longo alcance em estruturas teciduais complexas. Além disso, é possível estender a estrutura para *backbones* totalmente baseados em *transformers*, como DeiT, Swin Transformer e ViT-Base, a fim de compreender a eficácia da regularização sensível à entropia em modelos nativos de *self-attention*. Também é interessante observar o impacto das modificações propostas nos modelos no estado da arte para classificação de imagens histopatológicas, como o modelo DeepCMorph (Ignatov; Yates; Boeva, 2024). Por fim, para fortalecer as avaliações de generalização e interpretabilidade, outras frentes envolvem expandir o número de conjuntos de dados, explorando particularmente o Programa Atlas do Genoma do Câncer (TCGA), e

incorporar técnicas alternativas de explicação, como *attention rollout*, Score-CAM e métodos de saliência específicos para transformadores.

6.1 PUBLICAÇÕES

Durante o período de desenvolvimento deste trabalho também foram realizadas publicações em conferências e em revistas, sendo:

- Miguel, P.L.; Neves, L.A.; Lumini, A.; Medalha, G.C.; Roberto, G.F.; Rozendo, G.B.; Cansian, A.M.; Tosta, T.A.A.; do Nascimento, M.Z. "Improving Explainability of the Attention Branch Network with CAM Fostering Techniques in the Context of Histological Images". Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS. (Miguel. et al., 2024);
- Miguel, Pedro L., Leandro A. Neves, Alessandra Lumini, Giuliano C. Medalha, Guilherme F. Roberto, Guilherme B. Rozendo, Adriano M. Cansian, Thaína A. A. Tosta, and Marcelo Z. do Nascimento. 2025. "Entropy-Regularized Attention for Explainable Histological Classification with Convolutional and Hybrid Models"Entropy 27, no. 7: 722. (Miguel et al., 2025)

REFERÊNCIAS

- ABNAR, S.; ZUIDEMA, W. **Quantifying Attention Flow in Transformers**. 2020.
- AFFONSO, C. et al. Deep learning for biological image classification. **Expert Systems with Applications**, Elsevier, v. 85, p. 114–122, 2017.
- ALBUQUERQUE, T. et al. **Characterizing the Interpretability of Attention Maps in Digital Pathology**. 2024. Disponível em: <https://arxiv.org/abs/2407.02484>.
- ALOM, M. Z. et al. A state-of-the-art survey on deep learning theory and architectures. **Electronics**, v. 8, p. 292, 03 2019.
- Arrieta Legorburu, A. et al. Automatic discrimination between neuroendocrine carcinomas and grade 3 neuroendocrine tumors by deep learning of h&e images. **Computers in Biology and Medicine**, v. 184, p. 109443, 2025. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482524015282>.
- AYAZ, H. et al. An explainable deep learning model for multi-modal mri grading of idh-mutant astrocytomas. **Results in Engineering**, v. 24, p. 103353, 2024. ISSN 2590-1230. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2590123024016062>.
- BACH, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. **PLOS ONE**, Public Library of Science, v. 10, n. 7, p. 1–46, 07 2015. Disponível em: <https://doi.org/10.1371/journal.pone.0130140>.
- Baykal Kablan, E.; AYAS, S. Stainswin: Vision transformer-based stain normalization for histopathology image analysis. **Engineering Applications of Artificial Intelligence**, v. 133, p. 108136, 2024. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095219762400294X>.
- BELL, A. et al. It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: **Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2022. (FAccT '22), p. 248–266. ISBN 9781450393522. Disponível em: <https://doi.org/10.1145/3531146.3533090>.
- BI, Q. et al. Mil-vit: A multiple instance vision transformer for fundus image classification. **Journal of Visual Communication and Image Representation**, v. 97, p. 103956, 2023. ISSN 1047-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1047320323002067>.
- BROWN, T. B. et al. **Language Models are Few-Shot Learners**. 2020.
- BYLINSKII, Z. et al. **What do different evaluation metrics tell us about saliency models?** 2017. Disponível em: <https://arxiv.org/abs/1604.03605>.
- CHAU, M.; RAHMAN, M.; DEBNATH, T. From black box to clarity: Strategies for effective ai informed consent in healthcare. **Artificial Intelligence in Medicine**, p. 103169, 2025. ISSN 0933-3657. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0933365725001046>.

CORNIA, M. et al. Multi-level net: A visual saliency prediction model. In: HUA, G.; JÉGOU, H. (Ed.). **Computer Vision – ECCV 2016 Workshops**. Cham: Springer International Publishing, 2016. p. 302–315. ISBN 978-3-319-48881-3.

CORNIA, M. et al. Visual saliency for image captioning in new multimedia services. In: **2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)**. [S.l.: s.n.], 2017. p. 309–314.

DAI, Z. et al. **CoAtNet: Marrying Convolution and Attention for All Data Sizes**. 2021. Disponível em: <https://arxiv.org/abs/2106.04803>.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE Transactions on Evolutionary Computation**, v. 6, n. 2, p. 182–197, 2002.

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY. **XAI: Explainable Artificial Intelligence**. 675 North Randolph St., Arlington, VA 22203-1714, 2024. Disponível em: <https://www.darpa.mil/program/explainable-artificial-intelligence>. Acesso em: 08 nov. 2025.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DING, T. et al. A multimodal whole-slide foundation model for pathology. **Nature Medicine**, v. 31, n. 11, p. 3749–3761, nov 2025. ISSN 1546-170X. Disponível em: <https://doi.org/10.1038/s41591-025-03982-3>.

DOBBS, J. L. et al. Micro-anatomical quantitative optical imaging: toward automated assessment of breast tissues. **Breast Cancer Research**, v. 17, n. 1, p. 105, Aug 2015. ISSN 1465-542X. Disponível em: <https://doi.org/10.1186/s13058-015-0617-9>.

DONG, J. et al. Why did the ai make that decision? towards an explainable artificial intelligence (xai) for autonomous driving systems. **Transportation Research Part C: Emerging Technologies**, v. 156, p. 104358, 2023. ISSN 0968-090X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X23003480>.

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021.

Durand R., J. R. et al. Histattentionnas: A cnn built via nas for penile cancer diagnosis using histopathological images. **Procedia Computer Science**, v. 256, p. 764–771, 2025. ISSN 1877-0509. CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050925005344>.

DWIVEDI, K.; DUTTA, M. K.; PANDEY, J. P. Emvit-net: A novel transformer-based network utilizing cnn and multilayer perceptron for the classification of environmental microorganisms using microscopic images. **Ecological Informatics**, v. 79, p. 102451, 2024. ISSN 1574-9541. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1574954123004806>.

FILIPIAK, K. et al. Innovative methods of archiving, presentation and providing access to histological sections. **Medical Journal of Cell Biology**, v. 3, n. 3, p. 41–53, 2014. Disponível em: <https://doi.org/10.2478/v10052-011-0003-4>.

FISCHER, A. H. et al. Hematoxylin and eosin staining of tissue and cell sections. **CSH Protoc**, United States, v. 2008, p. db.prot4986, maio 2008.

FOMIN, V. et al. **High-level library to help with training neural networks in PyTorch**. [S.l.]: GitHub, 2020. <https://github.com/pytorch/ignite>.

FUKUI, H. et al. Attention branch network: Learning of attention mechanism for visual explanation. In: **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2019. p. 10697–10706.

GEEKS FOR GEEKS. **CNN: Introduction to pooling layer**. [S.l.], 2025. Disponível em: <https://www.geeksforgeeks.org/cnn-introduction-to-pooling-layer/>. Acesso em: 08 nov. 2025.

GELASCA, E. D. et al. Evaluation and benchmark for biological image segmentation. In: **2008 15th IEEE International Conference on Image Processing**. [S.l.: s.n.], 2008. p. 1816–1819.

GILDENBLAT, J. **Exploring Explainability for Vision Transformers**. São Paulo, 2023. Disponível em: <https://jacobgil.github.io/deeplearning/vision-transformer-explainability>. Acesso em: 08 nov. 2025.

GOCERI, E. Vision transformer based classification of gliomas from histopathological images. **Expert Systems with Applications**, v. 241, p. 122672, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417423031743>.

GOLSHANRAD, P.; FAGHIH, F. Deepcover: Advancing rnn test coverage and online error prediction using state machine extraction. **Journal of Systems and Software**, Elsevier BV, v. 211, p. 111987, maio 2024. ISSN 0164-1212. Disponível em: <http://dx.doi.org/10.1016/j.jss.2024.111987>.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing (3rd Edition)**. USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

GOODFELLOW, I. J. et al. **Generative Adversarial Networks**. 2014.

GRANDINI, M.; BAGLI, E.; VISANI, G. **Metrics for Multi-Class Classification: an Overview**. 2020. Disponível em: <https://arxiv.org/abs/2008.05756>.

GURINA, T. S.; SIMMS, L. **Histology, Staining**. [S.l.]: StatPearls Publishing, 2023.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HASSIJA, V. et al. Interpreting black-box models: A review on explainable artificial intelligence. **Cognitive Computation**, v. 16, n. 1, p. 45–74, Jan 2024. ISSN 1866-9964. Disponível em: <https://doi.org/10.1007/s12559-023-10179-8>.

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

HOCHREITER, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 6, p. 107–116, 04 1998.

HOSSIN, M. bin; M.N, S. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, v. 5, p. 01–11, 03 2015.

HUANG, G. et al. Densely connected convolutional networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Los Alamitos, CA, USA: IEEE Computer Society, 2017. p. 2261–2269. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.243>.

HUANG, Y. et al. **Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey**. 2024.

HÖHN, J. et al. Combining cnn-based histologic whole slide image analysis and patient data to improve skin cancer classification. **European Journal of Cancer**, v. 149, p. 94–101, 2021. ISSN 0959-8049. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959804921001404>.

HÖRST, F. et al. Cellvit: Vision transformers for precise cell segmentation and classification. **Medical Image Analysis**, v. 94, p. 103143, 2024. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841524000689>.

IGLESIAS, G.; MENENDEZ, H.; TALAVERA, E. Improving explanations for medical x-ray diagnosis combining variational autoencoders and adversarial machine learning. **Computers in Biology and Medicine**, v. 188, p. 109857, 2025. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482525002070>.

IGNATOV, A.; YATES, J.; BOEVA, V. **Histopathological Image Classification with Cell Morphology Aware Deep Neural Networks**. 2024. Disponível em: <https://arxiv.org/abs/2407.08625>.

IOANNIDIS, J. P. A.; MANIADIS, Z. In defense of quantitative metrics in researcher assessments. **PLoS Biol**, United States, v. 21, n. 12, p. e3002408, dez. 2023.

ISLAM, M. K. et al. Enhancing lung abnormalities diagnosis using hybrid dcnn-vit-gru model with explainable ai: A deep learning approach. **Image and Vision Computing**, v. 142, p. 104918, 2024. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885624000210>.

JAIN, S.; WALLACE, B. C. **Attention is not Explanation**. 2019. Disponível em: <https://arxiv.org/abs/1902.10186>.

JOHS, A. J.; AGOSTO, D. E.; WEBER, R. O. **Qualitative Investigation in Explainable Artificial Intelligence: A Bit More Insight from Social Science**. 2020.

KAMAKSHI, V.; KRISHNAN, N. C. Explainable image classification: The journey so far and the road ahead. **AI**, v. 4, n. 3, p. 620–651, 2023. ISSN 2673-2688. Disponível em: <https://www.mdpi.com/2673-2688/4/3/33>.

KANWAL, N. et al. Are you sure it's an artifact? artifact detection and uncertainty quantification in histological images. **Computerized Medical Imaging and Graphics**, v. 112, p. 102321, 2024. ISSN 0895-6111. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895611123001398>.

KASHEFI, R. et al. **Explainability of Vision Transformers: A Comprehensive Review and New Perspectives**. 2023.

KINDERMANS, P.-J. et al. **Learning how to explain neural networks: PatternNet and PatternAttribution**. 2017.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. **International Conference on Learning Representations**, 12 2014.

KINGMA, D. P.; WELLING, M. **Auto-Encoding Variational Bayes**. 2022. Disponível em: <https://arxiv.org/abs/1312.6114>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2012. v. 25. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

LEEM, S.; SEO, H. **Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention**. 2024.

LI, K. et al. **UniFormer: Unifying Convolution and Self-attention for Visual Recognition**. 2023. Disponível em: <https://arxiv.org/abs/2201.09450>.

LI, L. et al. An adaptive feature fusion framework of cnn and gnn for histopathology images classification. **Computers and Electrical Engineering**, v. 123, p. 110186, 2025. ISSN 0045-7906. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045790625001296>.

LI, X. et al. Improving feature extraction from histopathological images through a fine-tuning imagenet model. **Journal of Pathology Informatics**, v. 13, p. 100115, 2022. ISSN 2153-3539. Disponível em: <https://www.sciencedirect.com/science/article/pii/S215335392200709X>.

LI, Z. et al. A survey of convolutional neural networks: Analysis, applications, and prospects. **United States**, v. 33, n. 12, p. 6999–7019, nov. 2022.

LIU, S.; WANG, L.; YUE, W. An efficient medical image classification network based on multi-branch cnn, token grouping transformer and mixer mlp. **Applied Soft Computing**, v. 153, p. 111323, 2024. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494624000978>.

LIU, X.; HU, Y.; CHEN, J. Hybrid cnn-transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron. **Biomedical Signal Processing and Control**, v. 86, p. 105331, 2023. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809423007644>.

LIU, Z. et al. Deep learning face attributes in the wild. In: **Proceedings of International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2015.

LONGO, L. et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. **Information Fusion**, Elsevier BV, v. 106, p. 102301, jun. 2024. ISSN 1566-2535. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2024.102301>.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. 2017.

LÓPEZ, L. J. L. et al. **Uncertainty Quantification for Machine Learning in Healthcare: A Survey**. 2025. Disponível em: <https://arxiv.org/abs/2505.02874>.

Mahmud Kabir, S.; Imamul Hassan Bhuiyan, M. Cwc-mp-mc image-based breast tumor classification using an optimized vision transformer (vit). **Biomedical Signal Processing and Control**, v. 100, p. 106941, 2025. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809424009996>.

MAIER-HEIN, L. et al. Metrics reloaded: recommendations for image analysis validation. **Nature Methods**, Springer Science and Business Media LLC, v. 21, n. 2, p. 195–212, fev. 2024. ISSN 1548-7105. Disponível em: <http://dx.doi.org/10.1038/s41592-023-02151-z>.

MAJUMDAR, S.; PRAMANIK, P.; SARKAR, R. Gamma function based ensemble of cnn models for breast cancer detection in histopathology images. **Expert Systems with Applications**, v. 213, p. 119022, 2023. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417422020401>.

MAO, A.; MOHRI, M.; ZHONG, Y. Cross-entropy loss functions: theoretical analysis and applications. In: **Proceedings of the 40th International Conference on Machine Learning**. [S.l.]: JMLR.org, 2023. (ICML'23).

MARTINEZ, J. M. C. et al. Exploring deepdream and xai representations for classifying histological images. **SN Computer Science**, Springer, v. 5, n. 4, p. 362, 2024.

MIGUEL., P. et al. Improving explainability of the attention branch network with cam fostering techniques in the context of histological images. In: INSTICC. **Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS**. [S.l.]: SciTePress, 2024. p. 456–464. ISBN 978-989-758-692-7. ISSN 2184-4992.

MIGUEL, P. L. et al. Entropy-regularized attention for explainable histological classification with convolutional and hybrid models. **Entropy**, v. 27, n. 7, 2025. ISSN 1099-4300. Disponível em: <https://www.mdpi.com/1099-4300/27/7/722>.

MISRA, D. et al. Rotate to attend: Convolutional triplet attention module. In: **2021 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2021. p. 3138–3147.

MONTAVON, G. et al. Explaining nonlinear classification decisions with deep taylor decomposition. **Pattern Recognition**, Elsevier BV, v. 65, p. 211–222, maio 2017. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2016.11.008>.

NAMATĒVS, I. Deep convolutional neural networks: Structure, feature extraction and training. **Information Technology and Management Science**, Sciendo, v. 20, n. 1, p. 40–47, 2017.

OLIVEIRA, C. I. D. et al. Hybrid models for classifying histological images: An association of deep features by transfer learning with ensemble classifier. **Multimedia Tools and Applications**, Springer, v. 83, n. 8, p. 21929–21952, 2024.

PAN, X.-L. et al. El-cnn: An enhanced lightweight classification method for colorectal cancer histopathological images. **Biomedical Signal Processing and Control**, v. 100, p. 106933, 2025. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809424009911>.

PASZKE, A. et al. **PyTorch: An Imperative Style, High-Performance Deep Learning Library**. 2019. Disponível em: <https://arxiv.org/abs/1912.01703>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

POPPI, S. et al. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In: **2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2021. p. 2299–2304.

POWERS, D. M. W. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation**. 2020. Disponível em: <https://arxiv.org/abs/2010.16061>.

RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. In: . [s.n.], 2018. Disponível em: <https://api.semanticscholar.org/CorpusID:49313245>.

RADFORD, A. et al. **Language models are unsupervised multi-task learners**. 2019. Disponível em: <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>.

REZA, S.; AMIN, O. B.; HASHEM, M. Transresunet: Improving u-net architecture for robust lungs segmentation in chest x-rays. In: **2020 IEEE Region 10 Symposium (TENSYP)**. [S.l.: s.n.], 2020. p. 1592–1595.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**. 2016.

RICHE, N. et al. Saliency and human fixations: State-of-the-art and study of comparison metrics. In: **2013 IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2013. p. 1153–1160.

ROBERTO, G. F. et al. An ensemble of learned features and reshaping of fractal geometry-based descriptors for classification of histological images. **Pattern Analysis and Applications**, Springer, v. 27, n. 1, p. 8, 2024.

ROZENDO, G. B. et al. Data augmentation in histopathological classification: An analysis exploring gans with xai and vision transformers. **Applied Sciences**, MDPI, v. 14, n. 18, p. 8125, 2024.

ROZENDO, G. B. et al. Classification of non-hodgkin lymphomas based on sample entropy signatures. **Expert Systems with Applications**, Elsevier, v. 202, p. 117238, 2022.

RUSSAKOVSKY, O. et al. **ImageNet Large Scale Visual Recognition Challenge**. 2015.

SAEED, W.; OMLIN, C. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. **Knowledge-Based Systems**, v. 263, p. 110273, 2023. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705123000230>.

SANTOS, F. Pereira dos; PONTI, M. A. Alignment of local and global features from multiple layers of convolutional neural network for image classification. In: **2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2019. p. 241–248.

SARP, S. et al. An xai approach for covid-19 detection using transfer learning with x-ray images. **Heliyon**, v. 9, n. 4, p. e15137, 2023. ISSN 2405-8440. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405844023023447>.

SCHMIDT, R. M. **Recurrent Neural Networks (RNNs): A gentle Introduction and Overview**. 2019.

SCHÖTTL, A. Improving the interpretability of gradcams in deep classification networks. **Procedia Computer Science**, v. 200, p. 620–628, 2022. ISSN 1877-0509. 3rd International Conference on Industry 4.0 and Smart Manufacturing. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922002691>.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **2017 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 618–626.

SHAMIR, L. et al. Iicbu 2008: a proposed benchmark suite for biological image analysis. **Medical & Biological Engineering & Computing**, v. 46, n. 9, p. 943–947, Sep 2008. ISSN 1741-0444. Disponível em: <https://doi.org/10.1007/s11517-008-0380-5>.

SHIHABUDDIN, A. R.; K., S. B. Multi cnn based automatic detection of mitotic nuclei in breast histopathological images. **Computers in Biology and Medicine**, v. 158, p. 106815, 2023. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482523002809>.

SILVA, A. B. et al. Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue sections. **Expert Systems with Applications**, v. 193, p. 116456, 2022. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417421017401>.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**. 2014.

SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. 2015.

SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. 2015. Disponível em: <https://arxiv.org/abs/1409.1556>.

SINGH, A. et al. **Quantitative and Qualitative Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis**. 2021.

SIRINUKUNWATTANA, K. et al. Gland segmentation in colon histology images: The glas challenge contest. **Medical Image Analysis**, v. 35, p. 489–502, 2017. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841516301542>.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.

- SOOMRO, S.; NIAZ, A.; CHOI, K. N. Grad++scorecam: Enhancing visual explanations of deep convolutional networks using incremented gradient and score- weighted methods. **IEEE Access**, v. 12, p. 61104–61112, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:269347031>.
- SPRINGENBERG, J. T. et al. **Striving for Simplicity: The All Convolutional Net**. 2015.
- SZANDAŁA, T. Unlocking the black box of cnns: Visualising the decision-making process with prism. **Information Sciences**, v. 642, p. 119162, 2023. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025523007478>.
- TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. **ArXiv**, abs/1905.11946, 2019. Disponível em: <https://api.semanticscholar.org/CorpusID:167217261>.
- TANNER, E. M.; BORNEHAG, C.-G.; GENNINGS, C. Repeated holdout validation for weighted quantile sum regression. **MethodsX**, v. 6, p. 2855–2860, 2019. ISSN 2215-0161. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2215016119303103>.
- TAVANA EI, A. **Embedded Encoder-Decoder in Convolutional Networks Towards Explainable AI**. 2020.
- TENGUAM, J. J. et al. Ensemble learning-based solutions: An approach for evaluating multiple features in the context of h&e histological images. **Applied Sciences**, MDPI, v. 14, n. 3, p. 1084, 2024.
- THARWAT, A. Classification assessment methods. **Applied Computing and Informatics**, v. 17, n. 1, p. 168–192, 07 2020. ISSN 2634-1964. Disponível em: <https://doi.org/10.1016/j.aci.2018.08.003>.
- TOUVRON, H. et al. **Llama 2: Open Foundation and Fine-Tuned Chat Models**. 2023.
- TSUKAHARA, T. et al. Improving reliability of attention branch network by introducing uncertainty. In: IEEE. **2020 25th International Conference on Pattern Recognition (ICPR)**. [S.l.], 2021. p. 1536–1542.
- VASWANI, A. et al. **Attention Is All You Need**. 2023.
- VÁSQUEZ-VENEGAS, C. et al. Detecting and mitigating the clever hans effect in medical imaging: A scoping review. **Journal of Imaging Informatics in Medicine**, v. 38, 11 2024.
- WANG, Z.; POPEL, A.; SULAM, J. **CBM-zero: Concept Bottleneck Model With Zero Performance Loss**. 2024. Disponível em: <https://openreview.net/forum?id=wpL3otU9eY>.
- WOO, S. et al. **ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders**. 2023. Disponível em: <https://arxiv.org/abs/2301.00808>.
- XIE, S. et al. **Aggregated Residual Transformations for Deep Neural Networks**. 2017. Disponível em: <https://arxiv.org/abs/1611.05431>.
- XU, F. et al. Explainable ai: A brief survey on history, research areas, approaches and challenges. In: TANG, J. et al. (Ed.). **Natural Language Processing and Chinese Computing**. Cham: Springer International Publishing, 2019. p. 563–574. ISBN 978-3-030-32236-6.

YANG, S. et al. Bcd-net: Stain separation of histological images using deep variational bayesian blind color deconvolution. **Digital Signal Processing**, v. 145, p. 104318, 2024. ISSN 1051-2004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S105120042300413X>.

ZAHN, J. M. et al. Agedmap: A gene expression database for aging in mice. **PLOS Genetics**, Public Library of Science, v. 3, n. 11, p. 1–12, 11 2007. Disponível em: <https://doi.org/10.1371/journal.pgen.0030201>.

ZHOU, B. et al. **Object Detectors Emerge in Deep Scene CNNs**. 2015.

ZHOU, B. et al. Learning deep features for discriminative localization. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Los Alamitos, CA, USA: IEEE Computer Society, 2016. p. 2921–2929. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319>.

ZHOU, S. K.; GREENSPAN, H.; SHEN, D. **Deep learning for medical image analysis**. [S.l.]: Academic Press, 2017.

ZHUANG, F. et al. A comprehensive survey on transfer learning. **CoRR**, abs/1911.02685, 2019. Disponível em: <http://arxiv.org/abs/1911.02685>.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Pedro Lucas Miguel Data nascimento: 12/06/2001
Nacionalidade	Brasileiro
Nome em citações bibliográficas	MIGUEL, PEDRO LUCAS MIGUEL, P.L.
Currículo Lattes	http://lattes.cnpq.br/6884816922624752
ORCID	https://orcid.org/0009-0007-6809-2633
FORMAÇÃO ACADÊMICA	
2019/2023	Bacharel em Ciência da Computação Universidade Estadual Paulista "Júlio de Mesquita Filho"
2024/2026	Mestre em Ciência da Computação Universidade Estadual Paulista "Júlio de Mesquita Filho"
PRODUÇÃO BIBLIOGRÁFICA	
MIGUEL, P.; LUMINI, A.; CARDOZO MEDALHA, G.; FREIRE ROBERTO, G.; ROZENDO, G.; CANSIAN, A.; TOSTA, T.; NASCIMENTO, M. Z. do; NEVES, L. Improving Explainability of the Attention Branch Network with CAM Fostering Techniques in the Context of Histological Images. <i>In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS</i> , 26., 2024, Angers. Setúbal: SciTePress, 2024. p. 456-464.	
MIGUEL, P. L.; NEVES, L. A.; LUMINI, A.; MEDALHA, G. C.; ROBERTO, G. F.; ROZENDO, G. B.; CANSIAN, A. M.; TOSTA, T. A. A.; NASCIMENTO, M. Z. do. Entropy-Regularized Attention for Explainable Histological Classification with Convolutional and Hybrid Models. Entropy , Basileia, v. 27, n. 7, art. 722, 2025.	
PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES	
Bancas de trabalhos de conclusão	
Orientações	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	