

Otimização em Nível de Bloco da ResNet com Algoritmo Genético para Classificação de Imagens Endoscópicas

GILBERTO RODRIGUES DE SOUZA JUNIOR



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2026

Gilberto Rodrigues de Souza Junior

Otimização em Nível de Bloco da ResNet com
Algoritmo Genético para Classificação de Imagens
Endoscópicas

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do título de Mestre em Informática.

Orientador: Prof Dr. Lucídio dos Anjos Formiga Cabral

Fevereiro de 2026

Catálogo na publicação
Seção de Catalogação e Classificação

S729o Souza Junior, Gilberto Rodrigues de.
Otimização em nível de bloco da ResNet com algoritmo
genético para classificação de imagens endoscópicas /
Gilberto Rodrigues de Souza Junior. - João Pessoa,
2026.
83 f. : il.

Orientação: Lucídio dos Anjos Formiga Cabral.
Dissertação (Mestrado) - UFPB/CI.

1. Algoritmo genético. 2. Residual network (ResNet).
3. Redes neurais convolucionais (CNNs). 4.
Explicabilidade global. 5. Imagens endoscópicas. I.
Cabral, Lucídio dos Anjos Formiga. II. Título.

UFPB/BC

CDU 004.421(043)



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA



Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Gilberto Rodrigues de Souza Junior, candidato ao título de Mestre em Informática na área de Ciências da Computação, realizada em 11 de fevereiro de 2026.

1 Aos onze dias do mês de fevereiro do ano de dois mil e vinte e seis, às 11h, no auditório do
2 Centro de Informática, reuniram-se os membros da Banca Examinadora constituída para
3 julgar o Trabalho Final do discente **GILBERTO RODRIGUES DE SOUZA JUNIOR**, vinculado a
4 Universidade Federal da Paraíba sob a matrícula nº 20241005738, candidato ao grau de
5 Mestre em Informática, na área de “Ciências da Computação”, na linha de pesquisa
6 “Metodologia e Técnicas de Computação” do Programa de Pós-Graduação em Informática. A
7 comissão examinadora foi composta pelos professores Lucídio dos Anjos Formiga Cabral,
8 orientador e presidente da banca; Gilberto Farias de Sousa Filho, examinador interno ao
9 programa; Tiago Maritan Ugulino de Araujo, examinador interno ao programa; Roberto Quirino
10 do Nascimento, examinador externo ao programa; e Jefferson Queiroz Carneiro, Examinador
11 externo ao Programa. Dando início aos trabalhos, o presidente cumprimentou os presentes,
12 comunicou a finalidade da reunião e passou a palavra ao candidato para que fizesse,
13 oralmente, a exposição do trabalho de dissertação intitulado “**Otimização em Nível de Bloco**
14 **da ResNet com Algoritmo Genético para Classificação de Imagens Endoscópicas**”. Concluída
15 a exposição, o candidato foi arguido pela Banca Examinadora, que emitiu o seguinte parecer:
16 “**aprovado**”. Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário, lavrei a presente ata
17 que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 11 de
18 fevereiro de 2026.

Gean Paulo P. M. de Barros
Secretário - SIAPE 2326476

Prof. Dr. Lucídio dos Anjos Formiga Cabral
Orientador (PPGI)

Prof. Dr. Gilberto Farias de Sousa Filho
Examinador interno ao Programa (PPGI)

Prof. Dr. Tiago Maritan Ugulino de Araujo
Examinador interno ao Programa (PPGI)

Prof. Dr. Roberto Quirino do Nascimento
Examinador externo ao Programa (UFPB)

Me. Jefferson Queiroz Carneiro
Examinador externo ao Programa (HULW)

Documento assinado digitalmente
gov.br LUCIDIO DOS ANJOS FORMIGA CABRAL
Data: 19/02/2026 13:26:27-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente
gov.br GILBERTO FARIAS DE SOUSA FILHO
Data: 19/02/2026 16:02:04-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente
gov.br TIAGO MARITAN UGULINO DE ARAUJO
Data: 20/02/2026 14:12:07-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente
gov.br ROBERTO QUIRINO DO NASCIMENTO
Data: 19/02/2026 20:39:43-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente
gov.br JEFFERSON QUEIROZ CARNEIRO
Data: 20/02/2026 14:52:31-0300
Verifique em <https://validar.iti.gov.br>

AGRADECIMENTOS

Agradeço profundamente ao meu orientador, Prof. Dr. Lucídio dos Anjos Cabral, pela confiança depositada em meu potencial desde o início e pelos sábios conselhos ao longo da jornada. Ao Prof. Gilberto Farias, meu professor e revisor do artigo submetido, pela contribuição técnica e cuidadosa leitura crítica. Ao Prof. Dr. Tiago Maritan, por todo o aconselhamento especializado em deep learning, que foi essencial para o amadurecimento deste trabalho.

Sou imensamente grato à minha família: minha esposa Julianny, minha mãe Marlize e minha sogra Tânia, pelo suporte incondicional que me permitiu dedicar-me plenamente ao mestrado.

E, acima de tudo, agradeço ao Criador, fonte de sabedoria e propósito, por permitir que este caminho se abrisse diante de mim no tempo certo.

RESUMO

As Redes Neurais Convolucionais têm demonstrado capacidade na identificação de padrões em imagens médicas, embora a necessidade por modelos explicáveis permaneça crítica em aplicações médicas de alto risco. Este estudo apresenta um framework de otimização utilizando Neural Architecture Search com Algoritmo Genético para aprimorar o desempenho da arquitetura ResNet-152 e ResNet-50 no conjunto de dados Kvasir v2. A abordagem proposta incorpora técnicas de pré-processamento e otimização seletiva de blocos, alcançando acurácia de 97,75% no conjunto de validação e 97,12% no conjunto de teste. Adicionalmente, a análise do processo de otimização revelou indícios da relevância estrutural de blocos específicos de bottleneck, identificados como sensíveis às características do conjunto de dados. Os resultados deste estudo podem contribuir para o avanço da explicabilidade global em soluções baseadas na ResNet, oferecendo uma compreensão mais profunda do comportamento do modelo diante de imagens endoscópicas.

Relevância clínica — Este trabalho pode servir como base para a integração da interpretabilidade de modelos na imagem médica, possibilitando uma melhor compreensão das decisões baseadas na ResNet em exames endoscópicos, o que é essencial para sua aceitação clínica.

Palavras-chave: ResNet, Explicabilidade, Endoscopia, Algoritmo Genético, Kvasir

ABSTRACT

Convolutional Neural Networks have demonstrated capability in identifying patterns in medical images, however, the need for explainable models remains critical in high-risk medical applications. This study presents an optimization framework based on Neural Architecture Search using a Genetic Algorithm to improve the performance of ResNet-152 and ResNet-50 on the Kvasir v2 dataset. The proposed approach incorporates image preprocessing techniques and selective block optimization, achieving an accuracy of 97.75% on the validation set and 97.12% on the test set. In addition, the analysis of the optimization process revealed evidence of the structural relevance of specific bottleneck blocks, identified as being sensitive to dataset characteristics. The results of this study contribute to advancing global explainability in ResNet-based solutions by providing a deeper understanding of model behavior when applied to endoscopic images.

Clinical relevance — This work may serve as a foundation for integrating model interpretability into medical imaging, enabling improved understanding of ResNet-based decision-making in endoscopic examinations, which is essential for clinical acceptance.

Keywords: ResNet, Explainability, Endoscopy, Genetic Algorithm, Kvasir

LISTA DE FIGURAS

1	Fluxo de trabalho da solução proposta: YOLOv5 para pré-processamento, busca por algoritmo genético para padrões de congelamento em blocos, treinamento final com o melhor cromossomo, geração de mapas explicativos e auditoria usando métricas de localização e deleção.	40
2	Interface do programa LabelImg durante o processo de anotação manual. A imagem mostra uma caixa delimitadora sendo traçada sobre a imagem endoscópica, separando-a dos artefatos textuais.	48
3	Exemplo do nosso pré-processamento com YOLOv5. (a) Imagem endoscópica original contendo um exemplo da classe pólipos. (b) Caixa delimitadora detectada pelo YOLOv5 destacando a região do pólipo. (c) Imagem recortada com foco na ROI. Essa etapa reduz o ruído de fundo e direciona a CNN para as características mais relevantes.	49
4	Matriz de confusão da Resnet - 152 para o conjunto de teste, destacando o desempenho do modelo nas 8 classes: dyed-resection-margins, normal-cecum, normal-pylorus, normal-z-line, polyps, ulcerative-colitis, esophagitis e dyed-lifted-polyps. Os elementos da diagonal representam previsões corretas, onde a classe prevista coincide com a classe verdadeira, indicando a acurácia do modelo para cada categoria.	55
5	Matriz de confusão para o conjunto de dados de teste, destacando o desempenho da Resnet - 50 nas 8 classes: dyed-resection-margins, normal-cecum, normal-pylorus, linha Z normal, pólipos, ulcerative-colitis, esophagitis e dyed-lifted-polyps. Os elementos da diagonal representam previsões corretas, onde a classe prevista corresponde à classe verdadeira, indicando a precisão do modelo para cada categoria.	56
6	Curvas médias de deleção (Region Perturbation) para mapas Eigen-CAM por bloco, mostrando a evolução da probabilidade preditiva em função da fração removida. Os blocos finais (B13–B15) apresentam degradação mais precoce da probabilidade da classe alvo	58

7	Localização por classe ao longo dos blocos finais: ALIR para Eigen-CAM, reportando a média e IC95% para polyps, esophagitis e DRM nos blocos B10–B15. Observa-se aumento progressivo de ALIR em polyps e DRM nos blocos finais, enquanto esophagitis mantém valores baixos, refletindo menor concentração de relevância positiva dentro da ROI.	60
8	Convergência entre métodos de explicabilidade após regularização no espaço (ALIR, RP), considerando médias agregadas em todas as classes avaliadas. No modelo não regularizado, Grad-CAM e LRP ocupam regiões distintas, refletindo diferenças entre localização e impacto por deleção. Após a regularização, os pontos se aproximam, indicando redução da discrepância comportamental entre os métodos.	62
9	Comparação visual dos mapas LRP para esophagitis. Coluna esquerda: modelo não regularizado. Coluna direita: modelo regularizado. Primeira linha: overlay na imagem original. Segunda linha: heatmap isolado.	65
10	Comparação visual dos mapas LRP para pólipos. Coluna esquerda: modelo não regularizado. Coluna direita: modelo regularizado. Primeira linha: overlay na imagem original. Segunda linha: heatmap isolado.	66

LISTA DE TABELAS

1	Resumo dos estudos baseados em XAI em endoscopia	36
2	Trabalhos Relacionados na Classificação de Imagens Endoscópicas GI	37
3	Comparação de trabalhos por método XAI, validação e métricas de deleção.	38
4	Etapas do Algoritmo Genético	51
5	Blocos Treináveis: Frequências Absolutas e Relativas	54
6	Blocos Congelados: Frequências Absolutas e Relativas	54
7	Resultados de Treinamento para a Resnet - 152	56
8	Localização por Classe e Bloco: AUC de Deleção (Médias) e Métricas de Localização [IC 95%].	60
9	ALOR com LRP (R^-) Por classe [95% IC]	61
10	Resumo por classe e método: AUC de deleção [IC 95%] e localização (média).	63
11	Comparação de trabalhos por métrica, método de explicabilidade e modalidade.	64
12	Comparação de blocos no modelo regularizado: B15 - B12 (EigenCAM). Re- portado como Δ (diferença média), p_{Holm} , d_z e n	80
13	Localização no modelo não regularizado: AUC de deleção (médias) e métricas de localização [IC 95%].	81
14	Resumo por classe e método (modelo regularizado): média [IC 95%] para métricas de deleção e de localização.	82
15	Comparação Pareada GradCAM vs LRP (R^+): Δ , p_{Holm} , d_z , n	82

LISTA DE ABREVIATURAS

AUC	Area Under the Curve (Área sob a curva)
CAM	Class Activation Map
CNN	Convolutional Neural Network
GA	Genetic Algorithm (Algoritmo Genético)
GI	Gastrointestinal
IA	Inteligência Artificial
LIME	Local Interpretable Model-Agnostic Explanations
MF-CNN	Multi-Fusion Convolutional Neural Network
NAS	Neural Architecture Search
ROI	Region of Interest (Região de Interesse)
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
YOLO	You Only Look Once
XAI	Explainable Artificial Intelligence (Inteligência Artificial Explicável)
LRP	Propagação de Relevância por Camada
PG	Pointing Game
IROF	Iterative Removal Of Features
PF	Pixel Flipping
RP	Region Perturbation
ALIR	Localização da relevância positiva dentro da ROI
ALOR	Localização da relevância negativa fora da ROI
TKI	Top-K Intersection (Interseção dos K pixels mais relevantes com a ROI)
Grad-CAM	Gradient-weighted Class Activation Mapping
Eigen-CAM	Eigen Class Activation Mapping
SLIC	Simple Linear Iterative Clustering
ResNet	Residual Network
ViT	Vision Transformer
WCE	Wireless Capsule Endoscopy
CCE	Colon Capsule Endoscopy
BN	Batch Normalization
ReLU	Rectified Linear Unit
IC95	95% Confidence Interval (Intervalo de Confiança de 95%)
McNemar	McNemar's Test (Teste de McNemar)
Holm	Holm–Bonferroni Correction (Correção de Holm–Bonferroni)

Sumário

1	CAPÍTULO 1: INTRODUÇÃO	16
1.1	Justificativa	16
1.2	Panorama de trabalhos no domínio	17
1.3	Objetivo, contribuições e posicionamento	18
1.4	Objetivos	19
1.4.1	Objetivo Geral	19
1.4.2	Objetivos Específicos	19
1.5	Mapeamento das perguntas de pesquisa para os objetivos	20
1.6	Estrutura do Trabalho	20
2	CAPÍTULO 2: FUNDAMENTAÇÃO TEÓRICA	21
2.1	Redes Neurais Convolucionais (CNNs)	21
2.2	Importância da Explicabilidade em Saúde	22
2.3	Conceitos de Interpretabilidade em Modelos de Aprendizado de Máquina	23
2.4	Propagação de Relevância por Camada	24
2.5	Eigen-CAM: Visualização com Componentes Principais	24
2.6	Algoritmos Genéticos para Otimização	25
2.7	Pré-Processamento de Imagens com YOLOv5	27
2.8	Técnicas de Aumento de Dados (Data Augmentation)	28
2.9	Transfer Learning em Redes Neurais Convolucionais para Visão Computacional Médica	29
2.10	Métricas de Avaliação	29
3	CAPÍTULO 3: TRABALHOS RELACIONADOS	33
4	CAPÍTULO 4: METODOLOGIA	39
4.1	Solução Proposta	39
4.2	Conjunto de Dados KVASIR v2	40

4.3	Mapas de relevância via Grad-CAM	41
4.4	Eigen-CAM por bloco	41
4.5	Propagação de Relevância por Camada	42
4.6	Avaliação de Localização e Fidelidade Causal	42
4.6.1	Métricas de Localização	43
4.6.2	Métricas de Deleção	46
4.7	Configuração de Hardware	46
4.8	Pré-processamento com YOLOv5	47
4.9	Otimização Utilizando Algoritmo Genético	48
4.10	Treinamento Final	51
5	CAPÍTULO 5: ANÁLISE DOS RESULTADOS	53
5.1	Otimização com Algoritmo Genético	53
5.2	Treinamento Final	55
5.3	Explicação por Blocos	57
5.3.1	Testes de Deleção	57
5.3.2	Testes de Localização	59
5.4	Comparação de Métodos de Explicabilidade	61
5.4.1	Efeito da Regularização	61
5.4.2	Comparação com a literatura (ALIR/PG)	63
6	CAPÍTULO 6: CONCLUSÕES E TRABALHOS FUTUROS	67
6.1	DISCUSSÃO	67
6.2	CONCLUSÕES	68
	REFERÊNCIAS	69
	ANEXO A - ANEXOS E APÊNDICES 1	79

Nota sobre publicações associadas e reutilização de conteúdo

Esta dissertação consolida o mesmo projeto de pesquisa que originou duas versões públicas do trabalho. A primeira versão é o artigo: De Souza Junior, G. R., De Sousa Filho, G. F., & Cabral, L. D. A. F. (2025). *Block-Level Genetic Algorithm Optimization for ResNet Endoscopic Image Classification*. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, 1–5. doi:10.1109/EMBC58623.2025.11253109. A segunda versão é o preprint: *Which ResNet Blocks Matter and Why in Gastrointestinal Endoscopy*, depositado no TechRxiv e submetido ao *IEEE Transactions on Biomedical Engineering* (em avaliação). DOI: 10.36227/techrxiv.176591873.32954513/v1

Partes do texto, do método, dos resultados e de elementos gráficos foram adaptadas dessas versões, por tradução e reescrita para o português, podendo existir sobreposição parcial de redação.

1 CAPÍTULO 1: INTRODUÇÃO

A classificação de anomalias em exames endoscópicos gastrointestinais (GI) é uma área crescente na saúde, onde a aplicação de Redes Neurais Convolucionais (CNNs) tem sido amplamente adotada para aprimorar o diagnóstico médico. [Letafatkar et al. \[2025\]](#). A capacidade das CNNs de identificar padrões complexos em imagens médicas oferece uma oportunidade para aumentar a precisão diagnóstica [Litjens et al. \[2017\]](#).

No entanto, o uso de modelos de aprendizado profundo em áreas de alto risco, como a saúde, requer mais do que apenas alta precisão, se faz necessária uma justificativa clara das decisões tomadas, garantindo que as previsões possam ser verificadas por especialistas médicos [Guidotti et al. \[2018\]](#). Essa validação torna-se crítica, visto que as CNNs por exemplo, tendem a se basear em aprendizado por atalhos com base em correlações espúrias em vez de características patológicas reais, um fenômeno frequentemente associado ao efeito "Clever Hans" [Lapuschkin et al. \[2019\]](#).

Em aplicações médicas, explicar as previsões é essencial, pois impacta diretamente a segurança do paciente. A falta de interpretabilidade em modelos de aprendizado profundo, incluindo CNNs, que frequentemente são tratados como "caixas-pretas", representa um desafio para sua adoção em ambientes clínicos [Rudin \[2019\]](#).

1.1 Justificativa

Complementando esses desafios computacionais, a aceitação social também é um fator determinante para a implementação de técnicas de IA Explicável (XAI). O estudo de [Akinrinmade et al. \[2023\]](#) destaca que, devido à incapacidade dos algoritmos de aprendizado profundo de fornecer explicações compreensíveis sobre um diagnóstico, uma parcela dos usuários demonstra hesitação quanto ao uso de ferramentas de Inteligência Artificial em cenários críticos, ressaltando a necessidade de abordagens que mitiguem essa opacidade.

No contexto específico de imagens endoscópicas gastrointestinais, observamos um cenário em que estudos que apresentam alta acurácia geralmente não são acompanhados por abordagens de explicabilidade [Alshardan et al. \[2024\]](#); [Hossain et al. \[2024\]](#); [Khan et al. \[2022\]](#), enquanto estudos que abordam a explicabilidade fornecem principalmente explicações visuais baseadas apenas em mapas de saliência. Embora seja uma abordagem comum na literatura, ela é insuficiente para uma validação objetiva das previsões do modelo.

1.2 Panorama de trabalhos no domínio

Nesse sentido, [Khan et al. \[2022\]](#) introduziram o GestroNet, uma técnica baseada na estimativa de saliência profunda e otimização Bayesiana para a detecção de doenças gastrointestinais. Utilizando técnicas de aprimoramento de contraste, eles alcançaram acurácias entre 98,20% e 99,61% em diferentes conjuntos de dados, destacando os benefícios de estratégias avançadas de otimização em redes neurais.

[Alshardan et al. \[2024\]](#) apresentaram o algoritmo HSWOA-DLGCD para a detecção de câncer gastrointestinal, utilizando o XGBoost. Essa abordagem alcançou valores de precisão que variaram de 98,02% a 99,61% em três conjuntos de dados. Seguindo a mesma linha de desempenho baseado em ensembles, Hossain et al. desenvolveram o MF-CNN, uma arquitetura CNN de fusão múltipla que combina seis modelos distintos. Essa abordagem alcançou uma acurácia de 99,25% em testes com imagens endoscópicas [Hossain et al. \[2024\]](#).

Paralelamente, Jong et al. propuseram o GastroNet-5M, um modelo base auto-supervisionado pré-treinado em 4,8 milhões de imagens endoscópicas, e disponibilizaram os pesos do ResNet-50 treinados nesse corpus para facilitar aplicações de IA em endoscopia [Jong et al. \[2025\]](#); [Boers et al. \[2024\]](#). Esse desenvolvimento possibilita o aprimoramento de soluções anteriores baseadas em ResNet, substituindo o pré-treinamento genérico do ImageNet por aprendizado por transferência a partir de um conjunto de dados de endoscopia gastrointestinal.

Em relação a trabalhos sobre explicabilidade, Mukhtorov et al. propõem uma abordagem de aprendizado profundo explicável para imagens endoscópicas, combinando o Gradient Class Activation Mapping (Grad-CAM) com técnicas de aumento de dados. Essa combinação alcançou uma precisão de 93,46% e forneceu explicações visuais das regiões de interesse do modelo [Mukhtorov et al. \[2023\]](#).

De forma semelhante, Auzine et al. desenvolveram um conjunto de CNNs para classificação de câncer gastrointestinal. Usando as técnicas SHAP e LIME, os autores conseguiram identificar as regiões da imagem que mais influenciaram as previsões, as quais foram avaliadas por meio de inspeção visual [Auzine et al. \[2024\]](#).

No mesmo contexto de avaliação visual, Sutton et al. usam Kvasir v2 e HyperKvasir para duas tarefas binárias envolvendo colite ulcerativa, comparando qualitativamente mapas de calor com as anotações de um gastroenterologista em exemplos selecionados [Sutton et al. \[2022\]](#).

No trabalho de Nadimi et al., os autores propõem um pipeline de IA para Endoscopia

por Cápsula de Cólon (CCE) abrangendo detecção, caracterização e estimativa de tamanho de pólipos. O estudo combina múltiplas abordagens XAI, incluindo variantes de CAM e Relevância por Propagação em Camadas (LRP), para interpretar decisões de rede. A avaliação da explicabilidade é realizada por meio da inspeção visual dos mapas de saliência gerados, enquanto a avaliação quantitativa se limita a uma "pontuação única" não formalizada. [Nadimi et al. \[2025\]](#).

Incorporando aprendizado autossupervisionado na arquitetura de detecção, Zhao et al. desenvolveram o ETSEF, um conjunto que funde aprendizado por transferência supervisionado e aprendizado autossupervisionado contrastivo. A proposta de explicabilidade inclui Grad-CAM para localização, SHAP para contribuição de características e t-SNE para análise de distribuição latente [Zhao et al. \[2025\]](#).

O trabalho de Varam et al. adotou uma abordagem baseada em Transformer, onde os autores investigaram a classificação de imagens WCE usando um subconjunto balanceado de nove classes do conjunto de dados Kvasir-Capsule. O estudo comparou diversas arquiteturas de aprendizado profundo, destacando o desempenho do Vision Transformer (ViT), que alcançou uma pontuação F1 de aproximadamente 97%. Os autores integraram métodos baseados em perturbação (LIME, SHAP) e métodos baseados em gradiente (Grad-CAM, Grad-CAM++, LayerCAM). A avaliação da explicabilidade foi conduzida qualitativamente por meio da inspeção visual dos mapas de calor gerados, verificando se as regiões destacadas correspondiam a patologias clínicas [Varam et al. \[2023\]](#).

1.3 Objetivo, contribuições e posicionamento

Portanto, em resposta a esse cenário, o objetivo desta pesquisa é explicar o mecanismo de decisão do ResNet-50 para endoscopia gastrointestinal em nível de bloco. Além disso, buscamos descrever como as evidências pró-classe se concentram e analisar a distribuição dos contrassinais, aumentando assim a granularidade das explicações em comparação com os modelos descritos na literatura.

De forma complementar, em contraste com as avaliações visuais predominantemente qualitativas, nosso trabalho implementa métricas para medir a qualidade dos mapas de saliência, verificando sua precisão e regiões de foco. Também conduzimos uma análise objetiva para entender como componentes específicos da arquitetura do modelo processam imagens nesse domínio. A Tabela 1 resume como nossa abordagem de explicabilidade se relaciona com estudos anteriores de XAI em endoscopia gastrointestinal.

As principais contribuições deste trabalho consistem em aprimorar a transparência

dos mecanismos de decisão das CNNs em endoscopia gastrointestinal por meio de uma inspeção quantitativa utilizando três métodos de saliência (Grad-CAM, Eigen-CAM e LRP). Para tanto, implementamos técnicas de localização e análises de deleção para explicar objetivamente as decisões do modelo, fornecendo evidências empíricas que definem a confiabilidade em relação às limitações tanto do modelo avaliado quanto dos métodos de explicabilidade envolvidos.

1.4 Objetivos

1.4.1 Objetivo Geral

Alcançar acurácia comparável aos estudos mais relevantes em classificação de imagens endoscópicas para a base de dados proposta (Kvasir V2) [Pogorelov et al. \[2017\]](#), e também extrair explicações sobre o comportamento do modelo nesse domínio, visando entender a contribuição de blocos *bottleneck* e a relação destas ativações com achados clínicos.

1.4.2 Objetivos Específicos

1. **Quantificar a contribuição de blocos *bottleneck* da ResNet.**
Estimar a contribuição de cada bloco para a inferência por meio de testes de deleção e localização.
2. **Avaliar o impacto da remoção de artefatos no pré-processamento.**
Comparar resultados com e sem remoção de artefatos usando a métrica acurácia.
3. **Investigar o efeito do congelamento seletivo de blocos durante o fine-tuning.**
Testar estratégias para selecionar quais blocos treinar ou congelar, buscando aumentar a acurácia do modelo.
4. **Examinar correlações entre ativações de blocos e achados endoscópicos.**
Quantificar associações entre medidas de ativação por bloco, e achados clínicos usando métricas de deleção e localização para verificar quais blocos são causalmente mais relevantes para a rede.

1.5 Mapeamento das perguntas de pesquisa para os objetivos

Pergunta de pesquisa	Objetivo relacionado
Quais blocos <i>bottleneck</i> da ResNet apresentam contribuição estatisticamente significativa?	Objetivo 1
Em que medida a remoção de artefatos influencia a acurácia global?	Objetivo 2
É possível elevar a acurácia por meio do congelamento seletivo de blocos durante o fine-tuning?	Objetivo 3
Há correlação entre ativações de blocos específicos e determinados achados endoscópicos?	Objetivo 4

1.6 Estrutura do Trabalho

Esta seção descreve a organização lógica da dissertação e o papel de cada capítulo. O Capítulo 2 reúne os conceitos-chave necessários para entender o trabalho, incluindo a arquitetura da ResNet, a importância da explicabilidade em aplicações médicas, definições de interpretabilidade, técnicas como Eigen-CAM, transfer learning, algoritmos genéticos para otimização de blocos e hiperparâmetros, pré-processamento com YOLOv5 e técnicas de aumento de dados. Essa base teórica justifica as escolhas metodológicas detalhadas no Capítulo 4.

Em seguida, o Capítulo 3 apresenta e compara estudos relevantes sobre classificação e explicabilidade em imagens endoscópicas, identificando lacunas que este trabalho pretende preencher.

O Capítulo 4 descreve a solução proposta e os experimentos realizados, incluindo a especificação do conjunto Kvasir v2, a configuração de hardware e recursos computacionais, o pipeline de pré-processamento com detecção e recorte via YOLOv5, a estratégia de aumento de dados, a implementação do algoritmo genético para otimização do congelamento de blocos da ResNet, o protocolo de treino final, validação e testes, as métricas de avaliação e os procedimentos para análise das visualizações de ativação.

Na sequência, o Capítulo 5 apresenta os resultados quantitativos, como matrizes de confusão, métricas por classe e curvas de aprendizagem, e inclui uma discussão sobre robustez, generalização e comparação direta com os trabalhos selecionados no Capítulo 3.

Por fim, o Capítulo 6 resume os principais achados, avalia em que medida os objetivos foram atingidos e discute implicações clínicas e técnicas, propondo direções futuras.

2 CAPÍTULO 2: FUNDAMENTAÇÃO TEÓRICA

2.1 Redes Neurais Convolucionais (CNNs)

As Redes Neurais Convolucionais (CNNs) representam uma classe de redes neurais profundas que são particularmente eficazes no processamento de dados com estruturas em grade, como imagens, nas quais os pixels estão organizados de forma bidimensional. Conforme descrito por [LeCun et al. \[2015\]](#), as CNNs utilizam camadas convolucionais que aplicam filtros para detectar padrões locais, como bordas e texturas, em diferentes regiões da imagem. Cada camada convolucional é responsável por aprender características progressivamente mais complexas, partindo de detalhes simples, como linhas e bordas, até padrões mais avançados, como formas completas e objetos específicos.

[Litjens et al. \[2017\]](#) destacam que as Convolutional Neural Networks (CNNs) se tornaram amplamente utilizadas em tarefas de análise de imagens médicas, tornando-se a metodologia predominante devido à sua capacidade de aprendizagem direta e profunda a partir dos dados, sem necessidade de características manualmente projetadas. Essas redes são eficazes ao aplicar filtros convolucionais que capturam características em múltiplos níveis de abstração, desde padrões locais até representações mais complexas. Esse processo hierárquico de extração de características permite que as CNNs sejam eficazmente aplicadas em tarefas como a segmentação de órgãos, a classificação de lesões e a detecção de estruturas anômalas.

[He et al. \[2016\]](#) introduzem as Redes Residuais (ResNet) como uma solução inovadora para o problema de degradação do desempenho em redes neurais profundas, que ocorre quando o aumento da profundidade resulta em uma piora na acurácia. A abordagem proposta consiste em reformular o aprendizado de cada bloco de camadas para que estas aprendam uma "função residual", ao invés de diretamente tentar aprender a função desejada. Ou seja, as camadas são treinadas para modelar a diferença entre a saída desejada e a entrada do bloco, o que simplifica o processo de otimização e acelera o aprendizado. Essa estratégia possibilitou o treinamento eficaz de redes extremamente profundas, como a ResNet-152, que apresentou desempenho superior em comparação com redes tradicionais.

Além de estabelecer novos padrões de precisão no reconhecimento de imagens, a inovação das ResNets também se mostrou eficiente em outras tarefas de visão computacional, como detecção e segmentação de objetos, contribuindo significativamente para o avanço do campo.

2.2 Importância da Explicabilidade em Saúde

A explicabilidade em modelos de aprendizado de máquina tem ganhado uma importância crescente, especialmente quando aplicada em áreas críticas como a medicina. Um modelo explicável é aquele cujo funcionamento interno é compreensível para seres humanos, permitindo que suas decisões possam ser verificadas. Em aplicações médicas, a explicação das predições é essencial, pois envolve diretamente a segurança dos pacientes. A falta de interpretabilidade em modelos de deep learning, como as Redes Neurais Convolucionais (CNNs), que muitas vezes são tratados como "caixas-pretas", representa um grande desafio para a sua adoção em ambientes clínicos ([Rudin \[2019\]](#)).

Modelos "caixa-preta" são aqueles cujos processos de tomada de decisão são opacos, ou seja, não fornecem uma explicação clara de como as predições foram geradas, o que dificulta a interpretação por especialistas em saúde. Essa falta de transparência é especialmente preocupante em diagnósticos médicos, uma vez que decisões que afetam a vida dos pacientes precisam ser justificadas pelos profissionais de saúde para garantir a confiabilidade no processo. Como resultado, essa falta de clareza pode levar à hesitação por parte dos profissionais em adotar recomendações geradas por esses sistemas, especialmente se não houver uma base compreensível para a decisão. Por outro lado, um modelo explicável fornece uma visão clara dos fatores que contribuíram para uma determinada predição, permitindo que o especialista valide a decisão de maneira mais assertiva, além de verificar se ela está em conformidade com o conhecimento médico existente ([Guidotti et al. \[2018\]](#)).

A importância da interpretabilidade se reflete também na área de regulamentação. Muitas diretrizes, como o AI Act da União Europeia, exigem que sistemas de inteligência artificial sejam transparentes, principalmente em setores de alto risco como a medicina. Esses regulamentos visam garantir que as decisões geradas por sistemas automatizados possam ser verificadas e que, em caso de erro, seja possível rastrear as causas e corrigi-las ([Guidotti et al. \[2018\]](#)).

A interpretabilidade dos modelos pode ser classificada em duas categorias principais: explicações locais e explicações globais. As explicações locais são voltadas para entender como o modelo chegou a uma decisão específica, permitindo a análise da explicação de uma determinada predição ([Guidotti et al. \[2018\]](#)). Grad-CAM (Class Activation Mapping) é uma técnica amplamente utilizada para fornecer explicabilidade em modelos de redes neurais profundas, especialmente aqueles aplicados a dados visuais, como imagens médicas. Conforme destacado por [Guidotti et al. \[2018\]](#), o Grad-CAM utiliza uma combinação das ativações de camadas convolucionais e os gradientes de saída para gerar mapas de ativação que destacam

as regiões mais relevantes da imagem para a decisão do modelo. Esses mapas permitem que os médicos visualizem as áreas da imagem que foram mais determinantes para a classificação, tornando a predição do modelo mais transparente. (Guidotti et al. [2018]).

LIME (Local Interpretable Model-agnostic Explanations) é uma metodologia desenvolvida para fornecer explicações locais sobre as predições de modelos de machine learning, sendo agnóstica ao tipo de modelo utilizado. De acordo com Ribeiro et al. [2016], a técnica envolve a perturbação das entradas em torno da amostra específica a ser explicada e a criação de um modelo linear simples que aproxima o comportamento do modelo complexo na vizinhança da predição.

Esse modelo linear oferece uma interpretação visual da importância dos atributos, destacando quais características mais contribuíram para a predição do modelo. Dessa forma, LIME é particularmente útil em situações onde se deseja compreender decisões específicas e oferecer uma justificativa clara e simples para a predição de um exemplo individual (Guidotti et al. [2018]).

SHAP (SHapley Additive exPlanations), baseado em conceitos da teoria dos jogos, visa fornecer uma explicação quantitativa e visual do impacto de cada característica na predição do modelo. SHAP calcula os valores de Shapley para cada variável, determinando a contribuição individual de cada uma no resultado final do modelo. Essa técnica permite gerar gráficos de barras ou de dispersão que destacam a relevância de cada atributo, possibilitando uma visualização intuitiva de como cada característica afetou positivamente ou negativamente a predição. (Guidotti et al. [2018]).

Técnicas de explicabilidade pós-hoc, como o Grad-CAM, embora úteis, apresentam limitações importantes, pois não garantem uma fidelidade completa ao processo de decisão do modelo. Rudin [2019] argumenta que confiar exclusivamente nessas técnicas pode ser problemático, pois elas não necessariamente refletem com precisão o que o modelo realmente computou, levando a uma falsa sensação de confiança no modelo.

As explicações globais, por sua vez, são voltadas para oferecer uma visão geral sobre o comportamento do modelo como um todo, permitindo a compreensão de como os diferentes componentes do modelo interagem e quais são os principais padrões aprendidos durante o treinamento. (Guidotti et al. [2018]).

2.3 Conceitos de Interpretabilidade em Modelos de Aprendizado de Máquina

De acordo com Lipton [2017], o termo *interpretabilidade* não possui um significado técnico único na literatura. O autor propõe uma taxonomia conceitual que divide a inter-

pretabilidade em duas categorias principais: **transparência** e **explicações pós-hoc**.

A **transparência** refere-se à capacidade de compreender o funcionamento interno do modelo. Essa transparência pode ocorrer em diferentes níveis:

- *Simulatabilidade*: quando o modelo é suficientemente simples para ser compreendido por um ser humano em sua totalidade;
- *Decomposabilidade*: quando é possível interpretar individualmente cada entrada, parâmetro ou operação do modelo;
- *Transparência algorítmica*: quando o processo de treinamento do modelo é bem compreendido, previsível e reproduzível.

As **explicações pós-hoc** são geradas após o treinamento do modelo, com o objetivo de fornecer justificativas para suas decisões, não necessariamente revelando como o modelo funciona internamente, mas podem ser úteis para explicar as saídas do modelo. Alguns exemplos incluem mapas de saliência, visualizações de ativações, explicações por exemplos similares e modelos locais. [Lipton \[2017\]](#)

2.4 Propagação de Relevância por Camada

A Propagação de Relevância por Camada (Layer-wise Relevance Propagation, LRP) é uma técnica de explicabilidade que visa decompor a pontuação final produzida por uma rede neural em contribuições atribuídas aos componentes da entrada. No contexto de imagens, a LRP redistribui a evidência da decisão da rede ao longo das camadas, de forma conservativa, até o espaço de pixels, produzindo mapas de relevância que indicam quais regiões contribuíram positiva ou negativamente para a classe prevista. Bach et al. introduzem a LRP como um método para explicar decisões de modelos profundos por meio dessa decomposição da saída, permitindo distinguir entre evidências que apoiam e evidências que se opõem à classe prevista [Bach et al. \[2015\]](#).

Nesse contexto, Kohlbrenner et al. analisam diferentes variantes da LRP e propõem configurações que alcançam explicações mais estáveis [Kohlbrenner et al. \[2020\]](#).

2.5 Eigen-CAM: Visualização com Componentes Principais

A técnica Eigen-CAM, proposta por [Muhammad e Yeasin \[2020\]](#), é um método de explicabilidade que visa gerar mapas de ativação de classe (*Class Activation Maps* – CAMs) a

partir das representações aprendidas nas camadas convolucionais de redes neurais profundas, sem a necessidade de retropropagação de gradientes ou modificação da arquitetura do modelo.

Diferentemente de abordagens como Grad-CAM ou Grad-CAM++, que dependem de gradientes e pesos de relevância associados a uma classe específica, a Eigen-CAM realiza uma decomposição da saída da última camada convolucional utilizando *Singular Value Decomposition* (SVD). A técnica identifica os componentes principais da representação aprendida e utiliza o primeiro vetor singular (*first principal component*) como base para gerar um mapa de ativação discriminativo de classe.

O processo matemático pode ser resumido como:

$$O = WI, \quad (1)$$

onde I representa a imagem de entrada e W a matriz de pesos convolucionais. A seguir, aplica-se a decomposição:

$$O = U\Sigma V^T, \quad (2)$$

sendo U e V matrizes ortogonais, e Σ uma matriz diagonal com os valores singulares. O mapa de ativação da classe com o Eigen-CAM é obtido projetando a saída O no primeiro vetor singular:

$$L_{\text{Eigen-CAM}} = OV_1. \quad (3)$$

2.6 Algoritmos Genéticos para Otimização

[Elsken et al. \[2019\]](#) definem a Neural Architecture Search (NAS) como uma área que busca automatizar o design de redes neurais. Tradicionalmente, a criação de uma arquitetura de rede neural adequada era uma tarefa realizada manualmente por especialistas. A NAS procura resolver essas limitações através de algoritmos que possam, de forma automatizada, testar e ajustar diferentes arquiteturas para identificar aquelas que apresentam o melhor desempenho em uma tarefa específica.

Dentro do NAS, uma das abordagens mais promissoras para explorar arquiteturas inovadoras é a utilização de algoritmos genéticos. Segundo [Real et al. \[2019\]](#), esses algoritmos, inspirados pela evolução natural, constroem um espaço de busca que se assemelha a uma população de redes neurais, onde cada rede representa uma arquitetura única. Por meio de operações genéticas, como cruzamento e mutação, novas arquiteturas são iterativamente geradas e avaliadas, permitindo que as melhores estruturas “sobrevivam” e evoluam ao longo do tempo. Essa metodologia permite a otimização de CNNs de forma eficiente, ga-

mantendo diversidade nas arquiteturas e aumentando as chances de encontrar configurações bem adaptadas para problemas complexos.

A aplicação de algoritmos evolutivos na seleção de arquiteturas permite explorar um grande espaço de busca de modelos por meio de operações genéticas, como mutação e recombinação, gerando novas arquiteturas a partir de populações de redes previamente avaliadas.

Conforme descrito por [Elsken et al. \[2019\]](#), essa abordagem otimiza apenas a estrutura da rede, enquanto o treinamento dos pesos é realizado com métodos baseados em gradiente, como SGD, para garantir uma melhor performance no ajuste dos parâmetros de redes de grande escala. Além disso, os algoritmos genéticos no NAS permitem a evolução de populações de redes com diferentes arquiteturas. [Real et al. \[2019\]](#) demonstraram que, ao utilizar seleções por torneio e exclusões dos indivíduos mais antigos, é possível obter arquiteturas mais compactas e com desempenho robusto em comparação a métodos como busca aleatória e otimização baseada em aprendizado por reforço.

Segundo [Elsken et al. \[2019\]](#), algoritmos genéticos aplicados à busca por arquiteturas neurais utilizam principalmente três operações: seleção, mutação e, em alguns casos, crossover. A seleção consiste na escolha das arquiteturas que servirão como pais para gerar novas redes, geralmente com base em critérios de desempenho, como acurácia ou eficiência. A mutação é a principal operação usada, sendo definida como uma modificação local na arquitetura da rede, como adicionar ou remover camadas, alterar hiperparâmetros ou inserir conexões de atalho (skip connections).

O operador de *crossover*, conforme descrito por [Goldberg \[1989\]](#), é responsável por recombinar informações genéticas de dois indivíduos (pais) para gerar novos indivíduos (filhos), promovendo a exploração do espaço de busca. No algoritmo genético simples (SGA), o *crossover* é realizado com uma determinada probabilidade $p_{crossover}$. Caso ocorra, um ponto de corte j é escolhido aleatoriamente entre as posições 1 e $l - 1$, sendo l o comprimento do cromossomo.

Os genes anteriores a esse ponto são copiados diretamente dos pais correspondentes, enquanto os genes posteriores são trocados entre os pais, gerando dois filhos. Esse operador visa preservar e recombinar *blocos de construção* (building blocks) vantajosos, alinhando-se à hipótese de que boas soluções podem emergir da recombinação de subestruturas eficazes. Caso o *crossover* não ocorra, os filhos são cópias exatas dos pais, sujeitas apenas à aplicação do operador de mutação bit a bit.

2.7 Pré-Processamento de Imagens com YOLOv5

Artefatos em imagens médicas são sinais visuais que não correspondem à anatomia real do paciente, sendo introduzidos por limitações técnicas do equipamento, movimentação do paciente, interferência de objetos externos ou erros de reconstrução da imagem. Esses elementos podem simular patologias ou mascarar achados clínicos importantes, comprometendo a acurácia diagnóstica ([Nakashima e Duong \[2025\]](#)). No contexto endoscópico, artefatos comuns incluem saturação, baixo contraste, desfoque por movimento, reflexos especulares, bolhas, instrumentos, sangue e muco, conforme categorizado no desafio EndoCV2020 ([Ali et al. \[2021\]](#)).

Além disso, elementos extrínsecos como textos sobrepostos, timestamps ou indicadores gráficos do equipamento podem induzir viés nos modelos de deep learning e devem ser removidos ou mascarados no pré-processamento [KatherLab e Kather \[2022\]](#).

No contexto deste trabalho, o YOLOv5 será utilizado para realizar o pré-processamento das imagens de endoscopia, identificando para corte posterior, regiões relevantes da imagem, descartando elementos que não contribuem para a previsão, como textos, marcas de data e sinalizadores de posição do endoscópio. YOLO, que significa "You Only Look Once", é uma arquitetura desenvolvida para a detecção de objetos em tempo real ([Redmon e Farhadi \[2018\]](#)).

A principal inovação do YOLO é a capacidade de tratar a tarefa de detecção como um problema de regressão de ponta a ponta, em que uma única rede neural convolucional (CNN) é treinada para prever as caixas delimitadoras e as classes dos objetos simultaneamente a partir de uma única análise da imagem. Em comparação com abordagens anteriores, como Region-based Convolutional Neural Networks (R-CNN), o YOLO apresenta uma vantagem significativa em termos de velocidade, pois realiza todas as previsões em uma única etapa. Isso torna o YOLO particularmente adequado para aplicações que requerem processamento rápido, como análise de vídeo em tempo real ([Terven e Cordova-Esparza \[2023\]](#)).

O YOLOv4 e o YOLOv5 representaram avanços significativos na otimização e na eficiência do modelo. O YOLOv4 incorporou técnicas modernas, como bag of freebies (técnicas de aprimoramento de desempenho que não aumentam a complexidade do tempo de inferência, como o mosaic data augmentation) e bag of specials (estratégias como a Spatial Pyramid Pooling, que melhora a robustez da extração de características). O YOLOv5, amplamente utilizado pela comunidade, é valorizado por sua facilidade de uso, compatibilidade com bibliotecas modernas e melhorias em precisão e velocidade ([Terven e Cordova-Esparza \[2023\]](#)).

A arquitetura do YOLOv5 é baseada em uma versão modificada da CSPDarknet53, um backbone que combina a estrutura da Darknet, utilizada anteriormente no YOLOv3, com aprimoramentos conhecidos como Cross Stage Partial Connections (CSP). Esses aprimoramentos ajudam a reduzir a redundância de gradientes, melhorando o fluxo de informação durante o treinamento e aumentando a eficiência computacional (Terven e Cordova-Esparza [2023]).

Além do backbone aprimorado, o YOLOv5 também incorpora uma cabeça de detecção que realiza previsões em diferentes escalas, garantindo que o modelo possa detectar objetos de tamanhos variados dentro de uma imagem. A detecção em múltiplas escalas é feita através de uma estrutura com três cabeças de saída, cada uma ajustada para identificar objetos pequenos, médios e grandes, garantindo uma cobertura ampla das diferentes escalas presentes em uma imagem (Terven e Cordova-Esparza [2023]).

2.8 Técnicas de Aumento de Dados (Data Augmentation)

Data Augmentation, ou aumento de dados, é uma técnica essencial em aprendizado profundo, especialmente em aplicações que envolvem processamento de imagens, onde a quantidade de dados anotados pode ser limitada. Conforme discutido por Shorten e Khoshgoftaar [2019], o objetivo principal do Data Augmentation é gerar novas instâncias de dados a partir dos exemplos existentes, aumentando a diversidade dos dados de treinamento e, assim, melhorando a capacidade do modelo de generalizar para novos exemplos.

Overfitting é um problema comum em modelos de deep learning, onde a rede neural se torna excessivamente ajustada aos exemplos de treinamento, resultando em um mau desempenho em dados novos e desconhecidos. Data Augmentation desempenha um papel crucial na prevenção de overfitting ao aumentar a variabilidade do conjunto de treinamento. Ao aplicar transformações como rotações, espelhamentos, zoom, ajustes de brilho e ruídos, o modelo se torna mais robusto e menos propenso a memorizar padrões específicos dos dados originais. Essa variabilidade força o modelo a aprender representações mais gerais das características relevantes, o que aumenta a capacidade de generalização (Shorten e Khoshgoftaar [2019]).

Além da prevenção de overfitting, Data Augmentation é fundamental para aumentar o tamanho dos conjuntos de dados de treinamento, especialmente em contextos onde a coleta de dados é custosa ou demorada, como em aplicações médicas. Ao gerar mais amostras artificialmente, é possível fornecer mais dados para o treinamento, o que contribui diretamente para a melhora do desempenho do modelo, além de possibilitar um uso mais eficaz de redes

neurais profundas que precisam de grandes volumes de dados para aprender adequadamente (Shorten e Khoshgoftaar [2019]).

2.9 Transfer Learning em Redes Neurais Convolucionais para Visão Computacional Médica

O termo Transfer learning (aprendizado por transferência) refere-se a técnicas que aproveitam conhecimentos adquiridos em um domínio de origem para aprimorar o aprendizado em um domínio alvo relacionado. Essa abordagem permite reutilizar os pesos de um modelo previamente treinado em uma tarefa distinta ao invés de treinar um modelo inteiramente do início. Tal método inspira-se no comportamento humano onde habilidades aprendidas em um contexto podem ser aproveitadas para resolver problemas semelhantes em outro contexto (Pan e Yang [2010]).

O desempenho da AlexNet proposta por Krizhevsky et al. [2012] no desafio ImageNet em 2012, representou um marco no uso de redes neurais profundas. Sua arquitetura e pesos treinados se tornaram amplamente reutilizados em outras tarefas, estabelecendo as bases práticas para o uso difundido de transfer learning com CNNs.

Estudos posteriores demonstraram que as primeiras camadas das CNNs aprendem características visuais gerais, enquanto as últimas se especializam na tarefa original, afetando sua transferibilidade (Yosinski et al. [2014]). No domínio de imagens médicas, onde a escassez de dados anotados é comum, o transfer learning provou ser essencial para treinar modelos eficazes com conjuntos limitados Shin et al. [2016].

2.10 Métricas de Avaliação

As definições a seguir seguem a formalização proposta por Opitz [2024], que discute o uso dessas métricas em cenários com múltiplas classes.

Acurácia

A acurácia mede a proporção de instâncias corretamente classificadas em relação ao total de instâncias avaliadas:

$$accuracy = \frac{\sum_i m_{i,i}}{\sum_{(i,j)} m_{i,j}} = \frac{1}{|S|} \sum_i correct(i) \quad (4)$$

Onde:

- $m_{i,j}$: entrada da matriz de confusão na linha i e coluna j ;
- $m_{i,i}$: número de acertos para a classe i (elemento da diagonal da matriz de confusão);
- $|S|$: número total de instâncias no conjunto de dados;
- $correct(i) = m_{i,i}$: define o número de acertos na classe i .

Precisão Macro (Macro Precision)

A *precisão macro* é a média aritmética das precisões calculadas para cada classe:

$$macP = \frac{1}{n} \sum_i P_i = \frac{1}{n} \sum_i \frac{correct(i)}{bias(i)}. \quad (5)$$

Onde:

- P_i : precisão da classe i , definida como a razão entre acertos e o total de predições feitas para a classe i ;
- $correct(i)$: número de acertos na classe i , ou seja, $m_{i,i}$;
- $bias(i) = \sum_j m_{i,j}$: número total de predições feitas como sendo da classe i (linha i da matriz de confusão);
- n : número total de classes;
- $macP$: Macro Precision, ou seja, a média aritmética da precisão por classe.

Revocação Macro (Macro Recall)

A *revocação macro* é definida como a média da revocação por classe:

$$macR = \frac{1}{n} \sum_i R_i = \frac{1}{n} \sum_i \frac{correct(i)}{prevalence(i)}. \quad (6)$$

Onde:

- R_i : revocação (recall) da classe i , que representa a capacidade do modelo de identificar corretamente os exemplos dessa classe;

- $correct(i)$: número de acertos na classe i , ou seja, $m_{i,i}$;
- $prevalence(i) = \sum_j m_{j,i}$: total de instâncias reais da classe i (coluna i da matriz de confusão);
- n : número total de classes;
- $macR$: Macro Recall, média aritmética da revocação por classe, tratando todas as classes igualmente.

F1-Score Macro

O *F1-score macro* é a média dos F1-scores individuais de cada classe. A F1-Score Macro equilibra os efeitos de precisão e revocação em cada classe de forma equitativa:

$$macF1 = \frac{1}{n} \sum_i F1_i = \frac{1}{n} \sum_i \frac{2P_i R_i}{P_i + R_i} = \frac{2}{n} \sum_i \frac{correct(i)}{bias(i) + prevalence(i)}. \quad (7)$$

Onde:

- $F1_i$: F1-score da classe i , definido como a média harmônica entre P_i (precisão) e R_i (revocação);
- P_i : precisão da classe i , ou seja, $correct(i)/bias(i)$;
- R_i : revocação da classe i , ou seja, $correct(i)/prevalence(i)$;
- $correct(i) = m_{i,i}$: número de acertos da classe i ;
- $bias(i) = \sum_j m_{i,j}$: total de predições como sendo da classe i (linha da matriz de confusão);
- $prevalence(i) = \sum_j m_{j,i}$: total real da classe i (coluna da matriz de confusão);
- n : número total de classes;
- $macF1$: Macro F1, média dos F1-scores por classe.

Em síntese, este capítulo consolidou os conceitos necessários para sustentar a proposta desta dissertação: CNNs e redes residuais como base de modelagem, *transfer learning* como

estratégia de adaptação com dados limitados, pré-processamento com YOLOv5 para mitigar artefatos e viés, métricas para avaliar desempenho em múltiplas classes, e técnicas de explicabilidade, para inspecionar evidências utilizadas pelo modelo. Esses elementos definem o vocabulário técnico e os critérios de análise adotados ao longo do trabalho, permitindo discutir, nos capítulos seguintes, tanto o desempenho quanto a coerência das predições em imagens endoscópicas.

3 CAPÍTULO 3: TRABALHOS RELACIONADOS

Com a base teórica estabelecida, o próximo passo é situar esta pesquisa frente ao estado da arte em endoscopia gastrointestinal. Assim, o capítulo a seguir revisa estudos que aplicam CNNs e *transfer learning* em diferentes cenários, incluindo abordagens híbridas, arquiteturas especializadas e métodos de pré-processamento, e examina como a explicabilidade tem sido tratada, tipicamente por inspeção qualitativa e com pouca padronização quantitativa. Essa revisão permite identificar padrões metodológicos, métricas recorrentes e lacunas verificáveis, especialmente na validação de XAI e na análise de quais componentes da rede influenciam erros e acertos, delimitando o espaço em que a proposta desta dissertação se insere.

A combinação de CNNs com técnicas de aprendizado de máquina tradicionais, o desenvolvimento de novas arquiteturas, a aplicação de *transfer learning* e a preocupação com a interpretabilidade dos modelos têm contribuído para melhorias na acurácia dos modelos aplicados a exames de endoscopia digestiva.

Neste capítulo, é apresentada uma visão geral de trabalhos relacionados que exploram a classificação de lesões em endoscopia digestiva utilizando CNNs. A busca foi realizada nas bases de dados IEEE Xplore, PubMed e Scopus, empregando a string de pesquisa ("CNN"OR "deep learning") AND ("endoscopy"OR "gastrointestinal") AND ("diagnosis"OR "hybrid models"OR "transfer learning"OR "classification"OR "explainable"), com um recorte temporal de 2022 a 2024.

[Qin et al. \[2022\]](#), em seu trabalho, realizam uma revisão sistemática e meta-análise que demonstrou o desempenho das CNNs no diagnóstico de imagens de cápsula endoscópica. O estudo constatou uma sensibilidade variando entre 95% a 97% e especificidade entre 97% e 100%.

Há trabalhos na literatura que propuseram métodos híbridos que combinam CNNs com técnicas de aprendizado de máquina para aprimorar o diagnóstico. [Fati et al. \[2022\]](#) desenvolveram uma abordagem que utiliza GoogLeNet e AlexNet para extração de características, enquanto o SVM realiza a classificação, focando no diagnóstico precoce de doenças gastrointestinais (GI), obtendo uma Precisão 96,7% e AUC de 99,1%.

No trabalho de [Ahmed et al. \[2023\]](#), os autores empregaram metodologias híbridas que combinam CNNs com o algoritmo XGBoost, no conjunto de dados Kvasir. Essa combinação resultou em uma AUC de 97,54% e acurácia de 97,25% na classificação de imagens endoscópicas. Utilizando uma abordagem híbrida de otimização de redes neurais, [Khan et al. \[2022\]](#) introduziram o GestroNet, uma técnica baseada na estimativa de saliência profunda

e otimização Bayesiana para a detecção de GI.

O uso de técnicas de realce de contraste alcançou acurácia entre 98,20% e 99,61%, em diferentes bases de dados, evidenciando o benefício de otimizações avançadas em redes neurais. No artigo de [Alshardan et al. \[2024\]](#) é apresentado o algoritmo HSWOA-DLGCD para detecção de câncer gastrointestinal, que utiliza o XGBoost. Essa abordagem produziu valores de precisão entre 98,02% a 99,61% em 3 bases de dados.

Por outro lado, o aprimoramento de arquiteturas de rede tem sido o foco central. [Mary et al. \[2023\]](#), em sua pesquisa, desenvolveram o GastroNet, um modelo personalizado para a classificação de GI, alcançando 98,12% de acurácia. [Hossain et al. \[2024\]](#) desenvolveram o MF-CNN, uma arquitetura de CNN multi-fusão que integra seis modelos distintos. Essa abordagem alcançou uma acurácia de 99,25% em testes com imagens endoscópicas, demonstrando o potencial de modelos ensemble em capturar características complexas.

[Gunasekaran et al. \[2023\]](#) propuseram o GIT-Net, um modelo ensemble para classificação de GI, alcançando 95% de acurácia, sugerindo que a combinação de múltiplos modelos pode melhorar o desempenho diagnóstico. A transferência de aprendizado tem sido aplicada para melhorar a acurácia em tarefas diagnósticas.

[Sharma et al. \[2023\]](#) utilizaram o ResNet50 com transfer learning para o diagnóstico de GI em imagens endoscópicas, obtendo uma melhoria de 21%, em relação a outros modelos testados. Essa abordagem destaca como modelos pré-treinados em grandes conjuntos de dados podem ser adaptados eficazmente para tarefas específicas.

[Al-Otaibi et al. \[2024\]](#) apresentaram um modelo EfficientNet otimizado para a detecção de GI em imagens de cápsula endoscópica. O modelo atingiu 99,22% de precisão e uma AUC de 99,87%, no primeiro conjunto de dados, e uma AUC de 99,58%, no segundo conjunto de dados. Em paralelo, [Jong et al.](#) propuseram o GastroNet-5M, um modelo base auto-supervisionado pré-treinado em 4,8 milhões de imagens endoscópicas, e disponibilizaram os pesos do ResNet-50 treinados nesse corpus para facilitar aplicações de IA em endoscopia [Jong et al. \[2025\]](#); [Boers et al. \[2024\]](#).

Um segmento dos trabalhos na área de deep learning voltados para imagens de endoscopia, tem se concentrado em aumentar a explicabilidade dos modelos. [Mukhtorov et al.](#) propõem uma abordagem de aprendizado profundo explicável para imagens endoscópicas, combinando o Mapeamento de Ativação de Classe Gradiente (Grad-CAM) com técnicas de aumento de dados. Essa combinação alcançou uma acurácia de 93,46% e forneceu explicações visuais das regiões de interesse do modelo [Mukhtorov et al. \[2023\]](#).

De forma semelhante, [Auzine et al.](#) desenvolveram um conjunto de CNNs para clas-

sificação de câncer gastrointestinal. Utilizando as técnicas SHAP e LIME, os autores conseguiram identificar as regiões da imagem que mais influenciaram as previsões, as quais foram avaliadas por meio de inspeção visual [Auzine et al. \[2024\]](#).

No mesmo contexto de avaliação visual, Sutton et al. utilizam o Kvasir v2 e o HyperKvasir para duas tarefas binárias envolvendo colite ulcerativa, comparando qualitativamente mapas de calor com as anotações de um gastroenterologista em exemplos selecionados [Sutton et al. \[2022\]](#).

Nos trabalhos recentes em endoscopia, a explicabilidade aparece principalmente como verificação qualitativa, em que os autores inspecionam mapas de calor e julgam se a ativação coincide com regiões clinicamente plausíveis. Quando há quantificação, ela costuma ser indireta e pouco padronizada, o que limita comparações entre métodos.

Nadimi et al. enquadram um primeiro avanço nessa direção ao propor um pipeline para endoscopia por cápsula do cólon (CCE) que cobre detecção, caracterização e estimativa do tamanho de pólipos. Para interpretar as saídas da rede, o trabalho reúne variantes de CAM e Propagação de Relevância por Camadas (LRP). A análise é conduzida majoritariamente por inspeção visual dos mapas de saliência, e a parte numérica se resume a uma “pontuação única” sem definição formal, baseada na mudança de confiança do modelo após remover a região destacada [Nadimi et al. \[2025\]](#).

Zhao et al. seguem outra linha ao acoplar aprendizado autossupervisionado a uma arquitetura de detecção, por meio do ETSEF, que combina transferência supervisionada com aprendizado contrastivo autossupervisionado. Na explicabilidade, eles usam Grad-CAM para localização, SHAP para estimar contribuição de características e t-SNE para examinar a estrutura do espaço latente [Zhao et al. \[2025\]](#).

Varam et al. exploram uma abordagem com Transformer para classificar imagens de WCE, usando um subconjunto balanceado de nove classes do Kvasir-Capsule. O estudo compara arquiteturas e destaca o Vision Transformer (ViT), com F1 em torno de 97%. As explicações combinam métodos por perturbação, como LIME e SHAP, e métodos por gradiente, como Grad-CAM, Grad-CAM++ e LayerCAM. A validação de XAI permanece qualitativa, centrada na inspeção visual dos mapas e na checagem de correspondência com achados clínicos [Varam et al. \[2023\]](#).

O Quadro 2 sumariza as principais informações dos trabalhos relacionados a esta pesquisa. Enquanto que as tabelas 1 e 3 resumem como nossa abordagem de explicabilidade se relaciona com estudos anteriores de XAI em endoscopia gastrointestinal.

Ao consolidar os trabalhos analisados, observa-se que a literatura recente em endosco-

Tabela 1: Resumo dos estudos baseados em XAI em endoscopia

Author	XAI Escopo	Localização dos Mapas	Fidelidade Causal	Análise de Bloco
Sutton et al.	Local	Não	Não	Não
Nadimi et al.	Local	Não	Parcial	Não
Zhao et al.	Local	Não	Não	Não
Auzine et al.	Local	Não	Não	Não
Mukhtorov et al.	Local	Não	Não	Não
Varam et al.	Local	Não	Não	Não
Método Proposto	Global	Sim	Sim	Sim

pia gastrointestinal alcança métricas elevadas por meio de três eixos recorrentes: arquiteturas especializadas ou ensembles, *transfer learning* e pipelines de pré-processamento. Entretanto, permanecem lacunas para a robustez desses resultados.

Primeiramente, a explicabilidade é majoritariamente tratada como verificação qualitativa, com inspeção visual de mapas de calor e baixa padronização de métricas, o que limita comparações diretas entre métodos e dificulta quantificar fidelidade e consistência das explicações. Em segundo lugar, verificamos a ausência de metodologia para verificação de como componentes internos da rede, em especial blocos residuais, contribuem para as predições do modelo.

Diante dessas lacunas, este trabalho se posiciona ao investigar, em um protocolo reprodutível, o comportamento interno do modelo em endoscopia, combinando otimização em nível de blocos com análise de explicabilidade baseada em LRP, Grad-Cam e Eigen-CAM, e discutindo resultados à luz de métricas objetivas.

Tabela 2: Trabalhos Relacionados na Classificação de Imagens Endoscópicas GI

Autores	Objetivo	Métodos	Resultados
Fati et al. (2022)	Classificação de imagens endoscópicas do trato GI	Abordagem híbrida: GoogLeNet e AlexNet para extração de características; SVM para classificação	Precisão 96,7%; AUC de 99,1%
Ahmed et al. (2023)	Classificação de imagens endoscópicas do trato GI	Modelos híbridos combinando CNNs com o algoritmo XGBoost	AUC: 97,54%; Acurácia: 97,25%
Khan et al. (2022)	Classificação de imagens endoscópicas do trato GI	GestroNet: Estimativa de saliência profunda; otimização bayesiana; técnicas de realce de contraste	Acurácia: 98,20% a 99,61%
Alshardan et al. (2024)	Classificação de imagens endoscópicas do trato GI	Algoritmo HSWOA-DLGCD utilizando XGBoost; modelo híbrido com deep learning	Precisão: 98,02% a 99,61%
Mary et al. (2023)	Classificação de imagens endoscópicas do trato GI	GastroNet: Modelo de deep learning personalizado	Acurácia: 98,12%
Hossain et al. (2024)	Classificação de imagens endoscópicas do trato GI	MF-CNN: Arquitetura CNN multi-fusão integrando seis modelos	Acurácia: 99,25%
Gunasekaran et al. (2023)	Classificação de imagens endoscópicas do trato GI	GIT-Net: Modelo ensemble de deep learning; ponderação de modelos	Acurácia: 95%
Sharma et al. (2023)	Classificação de imagens endoscópicas do trato GI	ResNet50 com transferência de aprendizado	Melhoria de 21% em relação a outros modelos testados
Al-Otaibi et al. (2024)	Classificação de imagens endoscópicas do trato GI	Efficient-Gastro: Modelo EfficientNet otimizado com transferência de aprendizado	Precisão: 99,22%; AUC: 99,87% (primeiro conjunto); AUC: 99,58% (segundo conjunto)
Mukhtorov et al. (2023)	Classificação explicável de imagens endoscópicas	Deep learning explicável combinando Grad-CAM e aumento de dados	Acurácia: 93,46%; insights visuais sobre áreas de interesse do modelo
Auzine et al. (2024)	Classificação de câncer GI com IA explicável	Modelo ensemble de CNN; técnicas SHAP e LIME para explicabilidade	Identificação de regiões influentes nas previsões; melhor compreensão do modelo
Boers et al. (2024)	Classificação de imagens endoscópicas do trato GI	Pré-treinamento autossupervisionado em dados endoscópicos; avaliação em vários modelos	Melhoria de 2% a 6% na acurácia em três conjuntos de dados diferentes

Tabela 3: Comparação de trabalhos por método XAI, validação e métricas de deleção.

Trabalho	Método XAI	Validação	Métricas de Deleção
Sutton et al.	Grad-CAM	Inspeção visual	Não
Nadimi et al.	Grad-CAM, Deep SHAP	Inspeção visual	“Single Score”
Zhao et al.	Grad-CAM	Inspeção visual	Não
Mukhtorov et al.	Grad-CAM	Inspeção visual	Não
Varam et al.	Grad-CAM, SHAP, LIME	Inspeção visual	Não
Método proposto	Grad-CAM, LRP	PG, AL, TKI	PF, RP, IROF

4 CAPÍTULO 4: METODOLOGIA

Neste capítulo, descrevemos a metodologia adotada para implementar e avaliar a solução proposta, com ênfase em reprodutibilidade.

4.1 Solução Proposta

A solução proposta é composta por três etapas principais. Primeiramente, realiza-se o pré-processamento das imagens por meio de uma etapa de limpeza de dados (*data cleaning*) utilizando o modelo YOLOv5, com o objetivo de identificar e remover regiões irrelevantes ou artefatos visuais.

Em seguida, aplica-se um processo de otimização baseado em algoritmo genético para selecionar a configuração ideal de *transfer learning* da arquitetura ResNet-152, previamente carregada com os pesos da ImageNet. Essa etapa é conduzida sobre os conjuntos de treino e validação, visando identificar quais blocos residuais devem ser mantidos como treináveis e quais devem ser congelados. A configuração que resultar no melhor desempenho durante essa fase é então utilizada para realizar um treinamento final utilizando os conjuntos de treino, validação e teste.

Após a identificação da solução e consequente incremento na acurácia, foi conduzida uma análise da frequência de ativação dos blocos residuais mais frequentemente mantidos como treináveis nas melhores soluções.

Finalmente, geramos mapas explicativos (Grad-CAM, Eigen-CAM por blocos e LRP), avaliando-os com base em critérios de alinhamento espacial com as Regiões de Interesse (ROI) e fidelidade causal sob deleção.

Ressaltamos que a comparação da ROI mostrou-se mais adequada para classes com padrões de lesão focal, apresentando limitações para classes com padrões difusos. Nesses casos, a ROI cobriria uma região extensa e, portanto, as métricas de localização teriam menor poder conclusivo. Para fins de demonstração, mantivemos a classe de esofagite como uma exceção deliberada para destacar que, em padrões difusos, o modelo adota um comportamento distinto. Consequentemente, quando a lesão é difusa, nosso método permite formular hipóteses sobre o comportamento do modelo, no entanto, evitamos afirmações categóricas sobre o foco do modelo para essas classes (classes normais, Ulcerative-colitis e Esophagitis). A Figura 1 resume nosso fluxo de trabalho proposto.



Figura 1: Fluxo de trabalho da solução proposta: YOLOv5 para pré-processamento, busca por algoritmo genético para padrões de congelamento em blocos, treinamento final com o melhor cromossomo, geração de mapas explicativos e auditoria usando métricas de localização e deleção.

4.2 Conjunto de Dados KVASIR v2

O conjunto de dados KVASIR v2 é um banco público de imagens endoscópicas do trato gastrointestinal superior, desenvolvido com o objetivo de fomentar pesquisas em detecção automática de doenças gastrointestinais. Ele foi apresentado por [Pogorelov et al. \[2017\]](#) durante a 8ª Conferência ACM sobre Sistemas Multimídia (MMSys), e conforme verificável na Tabela 2, representa um dos benchmarks mais utilizados em tarefas de classificação em imagens endoscópicas.

A base contém um total de 8.000 imagens coloridas, organizadas em oito classes clínicas de igual tamanho (1.000 imagens por classe):

- **Dyed-lifted-polyps**(pólipos elevados com corante)
- **Dyed-resection-margins**(margens de ressecção com corante)
- **Esophagitis**(esofagite)
- **Normal-cecum**(ceco normal)
- **Normal-pylorus**(píloro normal)
- **Normal-z-line**(linha Z normal)
- **Polyps**(pólipos)
- **Ulcerative-colitis**(colite ulcerativa)

Cada imagem foi validada por médicos gastroenterologistas e está associada a uma única classe. As imagens foram extraídas de vídeos reais de exames endoscópicos realizados em pacientes adultos. As imagens variam em resolução de 720×576 a 1920×1072 pixels. O dataset inclui artefatos como reflexos especulares, presença de muco ou resíduos, textos sobrepostos e marcadores gráficos do equipamento endoscópico. ([Pogorelov et al. \[2017\]](#))

Este estudo utilizou imagens médicas anonimizadas, disponíveis publicamente, provenientes do conjunto de dados Kvasir-v2. Em conformidade com os princípios éticos estabelecidos na Declaração de Helsinque [World Medical Association \[2024\]](#), nenhum dado de identificação pessoal foi utilizado, e todo o material empregado no estudo provém de fontes públicas e anonimizadas.

4.3 Mapas de relevância via Grad-CAM

Para todos os métodos, adotamos o mesmo protocolo de pós-processamento e armazenamento de mapas: cada mapa bruto $\tilde{M} \in \mathbb{R}^{h \times w}$ é reamostrado via interpolação bilinear para a resolução nativa da imagem (H, W) , com o parâmetro `align_corners=false`. Em seguida, normalizamos cada amostra para $[0, 1]$ usando escala min-max com estabilização ε , da seguinte forma:

$$M = \frac{\tilde{M} - \min(\tilde{M})}{\max(\tilde{M}) - \min(\tilde{M}) + \varepsilon}$$

Salvamos M como float32 NPY [Vakanski et al. \[2020\]](#). Para inspeção visual, geramos sobreposições coloridas nas imagens originais, sem alterar as medições.

Geramos mapas Grad-CAM [Selvaraju et al. \[2017\]](#) a partir da última camada convolucional do último gargalo do ResNet-50 (layer4.2.conv2) [He et al. \[2016\]](#), usando a classe prevista como classe alvo.

4.4 Eigen-CAM por bloco

Para permitir uma comparação focada entre os blocos, aplicamos o Eigen-CAM [Muhammad e Yeasin \[2020\]](#) separadamente a cada um dos 16 blocos da rede. Dada uma imagem $I \in [0, 1]^{H \times W}$, primeiro realizamos uma passagem direta pela rede e, para cada bloco residual $b \in \mathcal{B}$, registramos os tensores de ativação $F^{(b)} \in \mathbb{R}^{C_b \times H_b \times W_b}$, onde C_b denota o número de canais e $H_b \times W_b$ a resolução espacial. Em seguida, aplicamos o Eigen-CAM bloco por bloco, produzindo mapas contínuos $M^{(b)}$ na grade da imagem. As etapas restantes, reamostragem, normalização e salvamento, seguiram o protocolo descrito anteriormente [Muhammad e Yeasin \[2020\]](#); [Pinciroli Vago et al. \[2021\]](#).

4.5 Propagação de Relevância por Camada

O método LRP [Bach et al. \[2015\]](#) decompõe o logit da classe alvo $f_c(x)$ em relevâncias R_i atribuídas aos pixels de entrada e neurônios internos. A partir do mapa de relevância LRP $R(x)$, separamos os componentes positivos e negativos como:

$$R^+(x) = \max(R(x), 0), \quad R^-(x) = \max(-R(x), 0),$$

interpretados respectivamente como evidência que apoia a classe alvo e evidência que contradiz a classe [Kohlbrenner et al. \[2020\]](#). Implementamos o LRP usando a biblioteca Zennit [Anders et al. \[2021\]](#); [Zennit Contributors \[2025\]](#), adotando o CompositeEpsilonGammaBox ($\varepsilon=10^{-4}$, $\gamma=0.25$) e o ResNetCanonizer para lidar com BatchNorm, ReLU e conexões de atalho dentro dos blocos. Os limites inferior e superior do domínio de entrada (“box”) foram definidos dentro do espaço normalizado do ImageNet. Essa configuração corresponde a uma variante composta do LRP (LRP_{CMP}), conforme discutido por [Kohlbrenner et al. \[2020\]](#).

4.6 Avaliação de Localização e Fidelidade Causal

No contexto da explicabilidade, localização refere-se ao alinhamento espacial entre o mapa de explicação e a ROI, enquanto fidelidade causal mede a extensão em que alterações em regiões destacadas produzem mudanças na saída do modelo [Hedström et al. \[2023\]](#). Neste trabalho, estruturamos a avaliação da fidelidade espacial em torno de três questões que investigam aspectos complementares do comportamento do classificador.

P1 Foco máximo do mapa de explicação: Os modelos podem alcançar alta precisão explorando correlações contextuais espúrias, como texto, marcas d’água ou artefatos visuais, em vez da descoberta clínica real. Lapuschkin et al. descrevem esse fenômeno de dependência sistemática de pistas irrelevantes como comportamento de "Clever Hans", associado a falhas de generalização fora do domínio de treinamento [Lapuschkin et al. \[2019\]](#). Para verificar se o modelo ancora sua decisão na lesão, utilizamos o Pointing Game, que contabiliza a frequência com que o pixel com a maior ativação se encontra dentro da ROI [Zhang et al. \[2018\]](#).

P2 Distribuição de evidências pró-classe entre lesão e contexto: Redes neurais profundas tendem a adotar atalhos, favorecendo correlações de textura e contexto que funcionam no conjunto de treinamento, mas reduzem a generalização [Geirhos et al. \[2020\]](#). Portanto, além de verificar o pico de ativação, é importante estimar qual fração do sinal da pró-classe está realmente concentrada na lesão. A métrica Attribution localization (AL) quantifica

a proporção de relevância positiva que se encontra dentro da ROI em relação à relevância total, permitindo a distinção entre decisões guiadas pela ROI e aquelas com foco difuso. [Kohlbrenner et al. \[2020\]](#).

P3 Cobertura da lesão pelas regiões mais salientes: Na localização fraca, sabe-se que os classificadores destacam apenas a parte mais discriminativa do objeto, por exemplo, o pico pode estar na lesão e a maior parte do que o modelo usou pode estar dentro da lesão, mas apenas uma pequena porção da lesão é coberta [Zhou et al. \[2015\]](#). Nesse sentido, a métrica Top- K Intersection (TKI) mede o quanto da região superior do mapa cobre a ROI [Theiner et al. \[2022\]](#). Assim, avaliamos a cobertura das regiões mais salientes sobre a lesão para responder quanta área da lesão foi efetivamente coberta pelos mapas.

Complementando a análise espacial, avaliamos o impacto causal dessas regiões na decisão do modelo. Ao usar mapas explicativos, os métodos de deleção servem para responder o que acontece durante a inferência do modelo quando regiões específicas são removidas. Samek et al. propuseram o Pixel Flipping (PF) [Samek et al. \[2017\]](#), que remove pixels em pequenos passos, sempre começando pelos mais relevantes no mapa explicativo. Se a precisão do modelo cair rapidamente à medida que esses pixels importantes são removidos, isso indica que o mapa está apontando para regiões que o modelo usa para sua decisão.

Petsiuk et al. refinaram essa ideia agrupando pixels em blocos regulares em um método chamado Region Perturbation (RP) [Petsiuk et al. \[2018\]](#), onde a imagem é dividida em patches quadrados e os patches mais importantes de acordo com o mapa são removidos primeiro.

Rieger et al. introduziram o Iterative Removal Of Features (IROF), uma métrica de deleção que avalia a fidelidade de um mapa de explicabilidade por meio da remoção iterativa de regiões da imagem [Rieger e Hansen \[2020\]](#). O IROF opera sobre superpixels obtidos por algoritmos de segmentação como o SLIC [Achanta et al. \[2012\]](#), de modo que as regiões removidas tendem a seguir contornos naturais e preservar estruturas locais. Em cada iteração, as regiões mais relevantes segundo o mapa de atribuição são removidas e observa-se a degradação da saída do modelo, permitindo quantificar o quanto a explicação captura evidências causalmente importantes para a predição.

4.6.1 Métricas de Localização

Para a análise de localização, quantificamos as seguintes métricas:

M1 Alinhamento de Pico (P1): A métrica Pointing Game verifica se o ponto de maior ativação do mapa está dentro da ROI. Definimos o Pointing Game (PG) para a amostra i como:

$$PG_i = \mathbb{I}\{\arg \max_p A_i(p) \in ROI_i\}.$$

Onde $A_i(p)$ denota o valor do mapa de atribuição no pixel p , $\arg \max_p A_i(p)$ representa a posição do valor máximo do mapa (pico), ROI_i é a máscara binária da lesão para essa imagem e $\mathbb{I}\{\cdot\}$ retorna 1 se a condição for verdadeira e 0 caso contrário. Em seguida, calculamos a PG média por classe em todo o conjunto de dados. A PG média representa a fração de imagens em que o pico do mapa se encontra dentro da lesão (por exemplo, 0,82 implica que, em 82% dos casos, o pico estava localizado dentro da ROI) [Zhang et al. \[2018\]](#).

M2 Precisão de Suporte da Pró-classe (P2): A métrica AL mede a fração de evidências positivas localizadas sobre a ROI. Neste trabalho, instanciamos AL em duas configurações: usando LRP (R^+) para a fração dentro da ROI e LRP (R^-) para a fração fora da ROI. Assim, estimamos a partir dos mapas LRP (R^+):

$$AL_i = \frac{\sum_p \max(A_i(p), 0) S_i(p)}{\sum_p \max(A_i(p), 0)}.$$

Onde $A_i(p)$ denota o valor do mapa de atribuição no pixel p . Utilizamos apenas valores positivos, $\max(A_i(p), 0)$, que representam evidências que apoiam a classe. $S_i(p) \in \{0, 1\}$ é a máscara binária da lesão, onde 1 corresponde ao interior da ROI e 0 ao exterior. O numerador soma as evidências positivas que se encontram dentro da ROI, enquanto o denominador soma o total de evidências positivas na imagem.

Após calcular AL_i por imagem, apresentamos a média por classe e a média global. Valores próximos de 1 indicam que a maior parte do suporte pró-classe está localizada dentro da lesão, valores baixos sugerem dependência de contexto, ou seja, suporte fora da ROI [Kohlbrenner et al. \[2020\]](#).

Para LRP (R^-), avaliamos a fração correspondente fora da ROI substituindo S_i por $(1 - S_i)$ e relatamos ambas as medidas. Para facilitar a compreensão, designamos essas métricas como Localização da Atribuição Dentro da ROI (ALIR), que quantifica a proporção de relevância positiva situada dentro da ROI, e Localização da Atribuição Fora da ROI (ALOR), que quantifica a proporção de relevância negativa fora da ROI. Essa interpretação é inspirada na separação bipolar de atribuições discutida por Nam et al. [Nam et al. \[2020\]](#).

M3 Cobertura Top-K (P3): A métrica de Interseção Top- K mede o quanto da parte superior do mapa cobre a lesão. Seja $\mathcal{I}_i^{(K)}$ o conjunto formado pelos K pixels com os maiores valores no mapa de atribuição A_i . Definimos:

$$\text{TKI}_i = \frac{|\mathcal{I}_i^{(K)} \cap \text{supp}(S_i)|}{K},$$

onde $S_i \in \{0, 1\}^{H \times W}$ é a máscara binária da lesão e $\text{supp}(S_i) = \{p : S_i(p) = 1\}$ é o conjunto de pixels dentro da ROI. Assim, o numerador conta quantos dos K pixels com maior valor estão dentro da lesão, e o denominador é o próprio K . Usamos $K = \lfloor \tau HW \rfloor$ com $\tau = 0, 15$, ou seja, selecionamos os 15% pixels com os valores mais altos no mapa.

Após calcular TKI_i por imagem, relatamos a média por classe e a média global. Valores altos indicam que as regiões mais salientes do mapa cobrem a lesão, enquanto valores baixos sugerem um ponto quente muito estreito ou um foco predominante fora da ROI [Theiner et al. \[2022\]](#).

As anotações da ROI foram realizadas por um estudante de medicina e validadas por um gastroenterologista especialista, utilizando a ferramenta LabelMe. Para cada imagem, a ROI foi definida por meio de um polígono fechado delineado diretamente ao redor da lesão visível, buscando abranger a área clinicamente relevante associada à classe diagnóstica correspondente.

O processo de anotação contemplou um subconjunto específico do dataset Kvasir v2, composto por quatro classes clínicas: esophagitis, polyps, dyed-lifted-polyps (DLP) e dyed-resection-margins (DRM). No total, foram anotadas 1 600 imagens do conjunto de treinamento, 400 imagens do conjunto de validação e 400 imagens do conjunto de teste, totalizando 2 400 imagens com ROIs em um dataset global de aproximadamente 8 000 imagens. As demais classes do dataset não foram incluídas no processo de anotação por não apresentarem, de forma consistente, lesões delimitáveis por ROI.

Após a etapa inicial de anotação, todas as ROIs foram submetidas à validação por um gastroenterologista especialista em endoscopia digestiva. A validação ocorreu com conhecimento prévio da classe associada a cada imagem, refletindo um cenário clínico supervisionado. O especialista avaliou a adequação espacial das ROIs em relação à lesão, assegurando que a área demarcada estivesse clinicamente coerente com o padrão patológico esperado para cada classe.

4.6.2 Métricas de Deleção

Nesta subseção, descrevemos a avaliação da fidelidade causal das explicações removendo progressivamente partes da imagem na ordem de importância indicada pelos mapas e observando a queda na probabilidade da classe alvo.

Para todos os métodos (PF, RP e IROF), adotamos o mesmo protocolo básico descrito abaixo: Usamos mapas $M \in \mathbb{R}^{H \times W}$ de Grad-CAM e Eigen-CAM, truncando valores negativos para zero. Para LRP, tratamos R^+ e a magnitude de R^- separadamente [Bach et al. \[2015\]](#); [Montavon et al. \[2017, 2019\]](#).

A ordem de remoção é decrescente com base na pontuação do mapa somada na unidade removida em cada iteração (pixel, patch ou superpixel). A única diferença entre os métodos é a granularidade dessa unidade; o critério de ordenação permanece o mesmo para todos.

Após selecionar a unidade com a maior pontuação, implementamos a remoção por perturbação, substituindo os valores da unidade selecionada por uma linha de base preta constante por canal, normalizada pela média/desvio padrão [Fong e Vedaldi \[2017\]](#); [Petsiuk et al. \[2018\]](#).

Para cada imagem e mapa, registramos a probabilidade da classe alvo na fração excluída e calculamos a AUC usando a regra trapezoidal. Como controle, geramos uma curva aleatória por imagem com a mesma granularidade e cronograma de etapas, embaralhando a ordem de remoção [Petsiuk et al. \[2018\]](#).

Agregamos AUCs e logits Δ por classe e globalmente (média, desvio padrão e n) [Samek et al. \[2017\]](#); [Petsiuk et al. \[2018\]](#); [Hedström et al. \[2023\]](#). Para PF, a unidade de remoção consiste em pixels individuais e o tamanho do passo é 3136 [Samek et al. \[2017\]](#); [Petsiuk et al. \[2018\]](#); [Fong e Vedaldi \[2017\]](#). Para testes RP, a unidade de remoção consiste em patches regulares $p \times p$ que cobrem completamente a imagem, com $p = 32$ para entradas 224×224 . Em IROF, a unidade de remoção consiste em superpixels obtidos via SLIC. Nesta etapa, usamos os seguintes parâmetros: `N_SEGMENTS = 0` e `COMPACTNESS = 10.0` [Rieger e Hansen \[2020\]](#); [Achanta et al. \[2012\]](#).

4.7 Configuração de Hardware

Todos os experimentos foram realizados em um ambiente local, utilizando um desktop equipado com uma GPU NVIDIA RTX 3060 com 12 GB de memória, processador AMD Ryzen 7 7800X, 64 GB de memória RAM DDR4 e armazenamento interno de 1 TB em SSD

NVMe.

4.8 Pré-processamento com YOLOv5

Para o pré-processamento das imagens, utilizamos o modelo YOLOv5 como etapa de limpeza e refinamento dos dados. O Kvasir-Dataset-v2 contém aproximadamente 8.000 imagens distribuídas em diferentes categorias clínicas. Para treinar a YOLOv5, selecionamos um subconjunto composto por 1.100 imagens anotadas manualmente com regiões de interesse (ROIs). Essa amostragem seguiu a seguinte divisão: 100 imagens para o Grupo 1 (Normal Z-line, Normal-Cecum, Normal-Pylorus); e 200 imagens para cada um dos demais grupos: Polyps, Ulcerative-Colitis, Dyed-Resection-Margins, Esophagitis e Dyed-Lifted-Polyps. Essa divisão foi escolhida pois observamos, empiricamente, que o desempenho da YOLOv5 era superior ao tratar as classes Normal Z-line, Normal-Cecum e Normal-Pylorus como um único grupo, dado o padrão visual semelhante entre elas e o caráter predominantemente normal das estruturas representadas.

Para a realização das anotações manuais, utilizamos a ferramenta LabelImg. Essa aplicação gráfica permite a marcação precisa das regiões de interesse (ROIs) nas imagens. A cada anotação realizada, o programa gera automaticamente um arquivo .txt no formato compatível com a YOLO, contendo as coordenadas normalizadas das caixas delimitadoras (bounding boxes) e o identificador da classe correspondente.

A Fig. 2 ilustra a interface gráfica do LabelImg com um exemplo de anotação manual sendo realizada, no qual se observa a separação entre artefatos textuais (por exemplo, informações sobre o paciente ou parâmetros técnicos exibidos na imagem) e a imagem endoscópica propriamente dita.

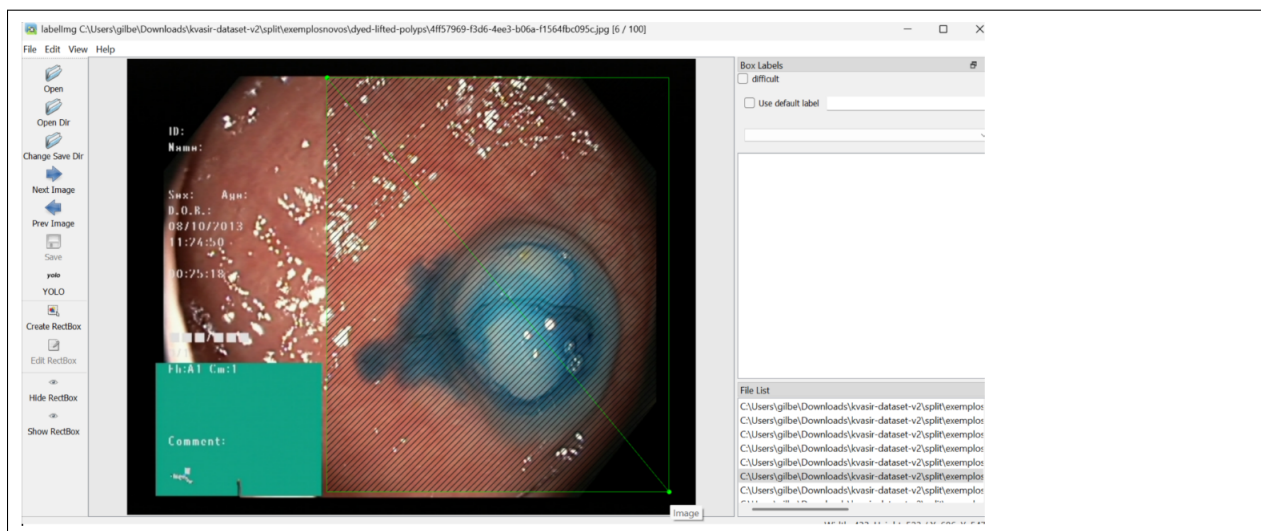


Figura 2: Interface do programa LabelImg durante o processo de anotação manual. A imagem mostra uma caixa delimitadora sendo traçada sobre a imagem endoscópica, separando-a dos artefatos textuais.

O conjunto anotado foi dividido em 90% para treinamento e 10% para validação, escolha motivada pela limitação de dados anotados e pela natureza objetiva da tarefa de detecção. Após o treinamento da YOLOv5, o modelo foi utilizado para realizar predições em todas as 8.000 imagens do dataset.

Na sequência, implementamos um script em Python para extrair as coordenadas das caixas delimitadoras previstas. Essas coordenadas foram então usadas para recortar as imagens, gerando uma nova versão do conjunto de dados com foco nas regiões de interesse identificadas automaticamente.

A Fig. 3 ilustra esse processo, evidenciando como o uso da YOLOv5 contribuiu para o pré-processamento.

4.9 Otimização Utilizando Algoritmo Genético

A implementação do modelo foi realizada em Python 3.8.20, com utilização da biblioteca PyTorch para construção e treinamento da arquitetura ResNet-152. [Paszke et al. \[2019\]](#). Para o carregamento do modelo pré-treinado, transformações de imagens e estruturação do conjunto de dados, foi utilizada a biblioteca [Torchvision maintainers e contributors \[2016\]](#). O processo evolutivo de seleção dos blocos *bottleneck* foi conduzido com a biblioteca DEAP (Distributed Evolutionary Algorithms in Python) [Fortin et al. \[2012\]](#). Para manipulação de vetores binários, geração de números aleatórios e análise estatística durante o

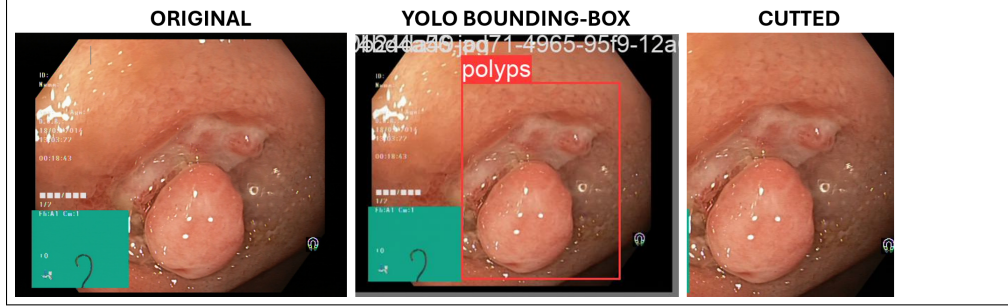


Figura 3: Exemplo do nosso pré-processamento com YOLOv5. (a) Imagem endoscópica original contendo um exemplo da classe pólip. (b) Caixa delimitadora detectada pelo YOLOv5 destacando a região do pólip. (c) Imagem recortada com foco na ROI. Essa etapa reduz o ruído de fundo e direciona a CNN para as características mais relevantes.

algoritmo genético, foi utilizada a biblioteca NumPy [Harris et al. \[2020\]](#). As visualizações dos resultados e métricas de desempenho foram realizadas com a biblioteca Matplotlib [Hunter \[2007\]](#).

Por padrão, a arquitetura ResNet no PyTorch permite o congelamento seletivo apenas no nível das camadas individuais dentro das camadas layer1 a layer4, não oferecendo suporte direto ao congelamento por bottleneck blocks. Para viabilizar o controle de congelamento por bloco, foi necessário iterar sobre os módulos internos dessas camadas e extrair diretamente cada bloco residual (Bottleneck) em uma lista ordenada. Essa estrutura permitiu aplicar configurações específicas de congelamento a cada bloco, manipulando diretamente o atributo `requires_grad` dos seus parâmetros, conforme definido por um vetor binário (cromossomo) gerado pelo algoritmo genético.

A otimização NAS do modelo ResNet-152, previamente carregado com pesos da ImageNet, foi realizada utilizando um Algoritmo Genético (GA), focando no congelamento e descongelamento seletivo dos blocos bottleneck, em vez de camadas individuais. Essa escolha se justifica pela diferença no tamanho do espaço de busca: tratar blocos como unidades de congelamento resulta em um número menor de configurações possíveis, reduzindo assim o custo computacional. Matematicamente, o espaço de busca associado à congelamento/descongelamento seletiva de blocos pode ser representado por variáveis binárias $(b_0, b_1, \dots, b_{N-1})$, onde $b_i = 1$ significa que o bloco i está congelado, e $b_i = 0$ indica que ele é treinável. Assim, o tamanho desse espaço de busca, $|\Omega|$, é dado por:

$$|\Omega| = 2^N$$

Para $N = 50$ blocos bottleneck, isso resulta em 2^{50} configurações, em comparação com

2^{152} se fossem consideradas camadas individuais. Formalmente, definimos um vetor binário $\mathbf{c} \in \{0, 1\}^N$ para descrever a configuração de congelamento dos N blocos bottleneck. Para cada bloco i ,

$$c_i = 0 \Rightarrow \text{treinável}, \quad c_i = 1 \Rightarrow \text{congelado}.$$

Os valores binários (0 ou 1) controlam quais partes da ResNet-152 são atualizadas durante o treinamento, permitindo a busca por uma solução sem a necessidade de explorar todo o espaço de busca.

A função de fitness no GA é definida como a acurácia de validação do modelo ResNet-152 configurado de acordo com um dado cromossomo. A avaliação envolve o treinamento do modelo com early stopping, que interrompe o treinamento quando a diferença entre a acurácia de treino e validação ultrapassa 3% após a quarta época. Parâmetros: número de épocas: 10, batch size: 24, learning rate: 1×10^{-5} , e random seed fixo: 3774091.

O conjunto de dados pré-processado foi dividido em 80% para treinamento, 10% para validação e 10% para teste. Durante a execução do GA foram utilizados apenas os dados de treino e validação. Além disso, a estratégia de data augmentation aplicada às imagens de entrada combinou redimensionamento, rotações aleatórias (15°), espelhamento horizontal aleatório, com probabilidade de 50% de aplicar as transformações aleatórias, seguidas de normalização aplicada a todas as imagens.

Para a Resnet - 152, executamos o algoritmo por 10 gerações e percebemos que alguns blocos estavam consistentemente congelados nos melhores resultados, então decidimos repetir o experimento impondo uma restrição, definindo esses blocos como permanentemente congelados: (1,2,6,7,13,16,18,22,26,28,29,30,32,33,35,37,42,49). Os demais parâmetros do GA são detalhados na Tabela 4. Para a Resnet - 50, devido ao tamanho reduzido do modelo, não aplicamos essa abordagem de congelamento permanente de alguns blocos.

Tabela 4: Etapas do Algoritmo Genético

Etapa	Descrição
Cromossomo	$\mathbf{c} \in \{0, 1\}^N$: $c_i = 0$ (treinável), $c_i = 1$ (congelado).
Congelamento	FreezeModel($\mathbf{c}$): Configura os parâmetros congelados.
População	$P^g = \{\mathbf{c}_1, \dots, \mathbf{c}_M\}$: Fitness é a acurácia de validação.
Seleção	Pais selecionados por torneio (tamanho 3).
Crossover	Crossover de dois pontos gera os descendentes.
Mutação	Mutação do tipo bit-flip com probabilidade $p_{\text{mut}}=0.05$.
Próxima Geração	$P^{g+1} = \text{Replacement}(P^g, O^g)$.
Critério de Parada	Parada após 50 gerações.

4.10 Treinamento Final

O treinamento final foi conduzido utilizando o cromossomo que retornou a maior acurácia durante o GA, que determina quais blocos bottleneck da ResNet-152 permanecerão treináveis ou congelados. O modelo, com pesos pré-carregados da ImageNet, foi treinado por 15 épocas com uma estratégia dinâmica de taxa de aprendizado.

Na primeira época, utilizou-se uma taxa de aprendizado mais alta (1×10^{-4}) para acelerar a convergência inicial. A partir da segunda época, a taxa de aprendizado alternou entre 1×10^{-5} (épocas pares) e 1×10^{-6} (épocas ímpares). Durante a etapa de treinamento final, foram utilizados os subconjuntos de treino e validação para otimização dos hiperparâmetros e ajuste final do modelo.

Após essa fase, o conjunto de teste foi utilizado exclusivamente para a geração das previsões finais, sem qualquer intervenção adicional no modelo. A partir desse ponto, o modelo permaneceu inalterado para garantir a integridade da avaliação final.

Para a Resnet - 50 aplicamos uma etapa de regularização estocástica envolvendo a introdução de uma camada Dropout ($p = 0,3$) precedendo o cabeçalho do classificador final [Srivastava et al. \[2014\]](#), e a adoção de uma função de perda composta que combina a Entropia Cruzada com três termos auxiliares: uma penalidade de borda para desencorajar ativações de borda, uma recompensa ROI para incentivar o foco na lesão e uma penalidade de artefato

para suprimir ativações em máscaras de artefatos [Rieger et al. \[2020\]](#).

Em síntese, a metodologia estabelece um protocolo completo para avaliar, de forma integrada, desempenho e comportamento interno do classificador em endoscopia.

A combinação de pré-processamento com YOLOv5, otimização por algoritmo genético no nível de blocos e treinamento final define a configuração experimental usada para reportar acurácia e métricas macro no conjunto Kvasir-v2.

Em paralelo, a geração de mapas por Grad-CAM, Eigen-CAM por bloco e LRP, seguida da auditoria por métricas de localização e deleção, fornece evidência mensurável sobre alinhamento com a ROI e impacto causal das regiões destacadas na decisão do modelo. No capítulo seguinte, apresentamos os resultados quantitativos e qualitativos desse protocolo.

5 CAPÍTULO 5: ANÁLISE DOS RESULTADOS

Para validar o método proposto, realizamos experimentos com o objetivo de otimizar o desempenho e auditar a explicabilidade do modelo. Primeiro, executamos o Algoritmo Genético e analisamos as configurações que alcançaram a maior precisão de validação. Posteriormente, treinamos o modelo final usando o melhor cromossomo e relatamos os resultados no conjunto de teste Kvasir-v2. Para medir a generalização fora do domínio, aplicamos este mesmo modelo, sem ajuste fino adicional, aos conjuntos de dados CVC-ClinicDB e Hyper-Kvasir-SEG [Bernal et al. \[2017\]](#); [Borgli et al. \[2020\]](#). Finalmente, avaliamos as explicações por bloco e comparamos dois métodos de explicabilidade usando métricas de localização e deleção.

5.1 Otimização com Algoritmo Genético

Após o processo de otimização, para a Resnet - 152 obtivemos um total de 3363 configurações, dessas, 105 alcançaram acurácia igual ou superior a 97 %. Durante os experimentos, verificamos que nosso método de pré-processamento contribuiu com um incremento de aproximadamente 2% na acurácia, enquanto a estratégia de otimização baseada em algoritmo genético adicionou mais 2% de acurácia ao modelo. A análise de frequência dos blocos treináveis e congelados revelou padrões na configuração dos blocos *bottleneck* entre as soluções avaliadas.

Por exemplo, os blocos 47 e 45 foram definidos como treináveis em 101 das 105 melhores soluções. A Tabela 5 apresenta as frequências absolutas e relativas dos 10 blocos mais frequentemente configurados como treináveis. Enquanto que A Tabela 6 apresenta as frequências absolutas e relativas dos 10 blocos mais frequentemente configurados como congelados.

Em relação aos resultados de otimização na RESNET - 50, obtivemos um total de 1.023 configurações, das quais 68 alcançaram uma precisão de 98% ou superior no conjunto de validação. A análise de frequência de blocos treináveis e congelados revelou padrões na configuração dos blocos de gargalo entre as soluções avaliadas. Por exemplo, os blocos 4 e 9 foram definidos como treináveis 54 e 53 vezes, respectivamente, entre as 68 melhores soluções.

Tabela 5: Blocos Treináveis: Frequências Absolutas e Relativas

Bloco	Frequência Absoluta	Frequência Relativa
47	101	0,9619
45	101	0,9619
19	99	0,9429
24	98	0,9333
0	96	0,9143
25	96	0,9143
27	96	0,9143
38	96	0,9143
9	95	0,9048
36	95	0,9048

Tabela 6: Blocos Congelados: Frequências Absolutas e Relativas

Bloco	Frequência Absoluta	Frequência Relativa
7	105	1,0
22	105	1,0
28	105	1,0
30	105	1,0
32	105	1,0
33	105	1,0
35	105	1,0
37	105	1,0
49	105	1,0
26	104	0,9905

5.2 Treinamento Final

A fase final de treinamento utilizou o cromossomo que alcançou a maior acurácia de validação, atingindo uma acurácia de 0,9712 no conjunto de dados de teste para ambas as redes.

Na ResNet-152, as classes dyed-lifted-polyps (DLP), esophagitis e normal-z-line apresentaram acurácia abaixo de 97%, enquanto as demais classes permaneceram acima de 97%, conforme a matriz de confusão da Figura 4.

Na ResNet-50 regularizada, a matriz de confusão da Figura 5 indica acurácia acima de 97% para esophagitis, normal-z-line, normal-pylorus, polyps e ulcerative-colitis (UC). Em contraste, dyed-resection-margins (DRM), normal-cecum e (DLP) ficaram abaixo de 97%. No modelo não regularizado, embora a acurácia global tenha sido 97,62%, observamos degradação nas métricas de localização e deleção, aspecto analisado mais adiante.

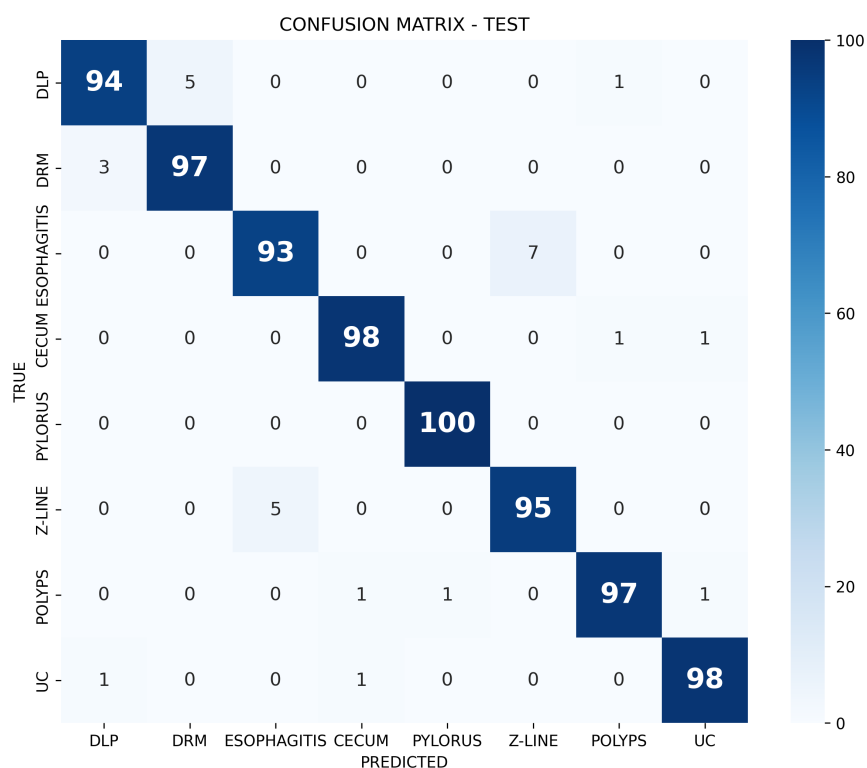


Figura 4: Matriz de confusão da Resnet - 152 para o conjunto de teste, destacando o desempenho do modelo nas 8 classes: dyed-resection-margins, normal-cecum, normal-pylorus, normal-z-line, polyps, ulcerative-colitis, esophagitis e dyed-lifted-polyps. Os elementos da diagonal representam predições corretas, onde a classe prevista coincide com a classe verdadeira, indicando a acurácia do modelo para cada categoria.

Tabela 7: Resultados de Treinamento para a Resnet - 152

Métrica	Treinamento	Validação	Teste
Acurácia	0.9936	0.9775	0.9712
Precisão (Macro)	-	0.9777	0.9716
Revocação (Macro)	-	0.9775	0.9712
F1-Score (Macro)	-	0.9775	0.9712
Precisão (Ponderada)	-	0.9777	0.9716
Revocação (Ponderada)	-	0.9775	0.9712
F1-Score (Ponderado)	-	0.9775	0.9712

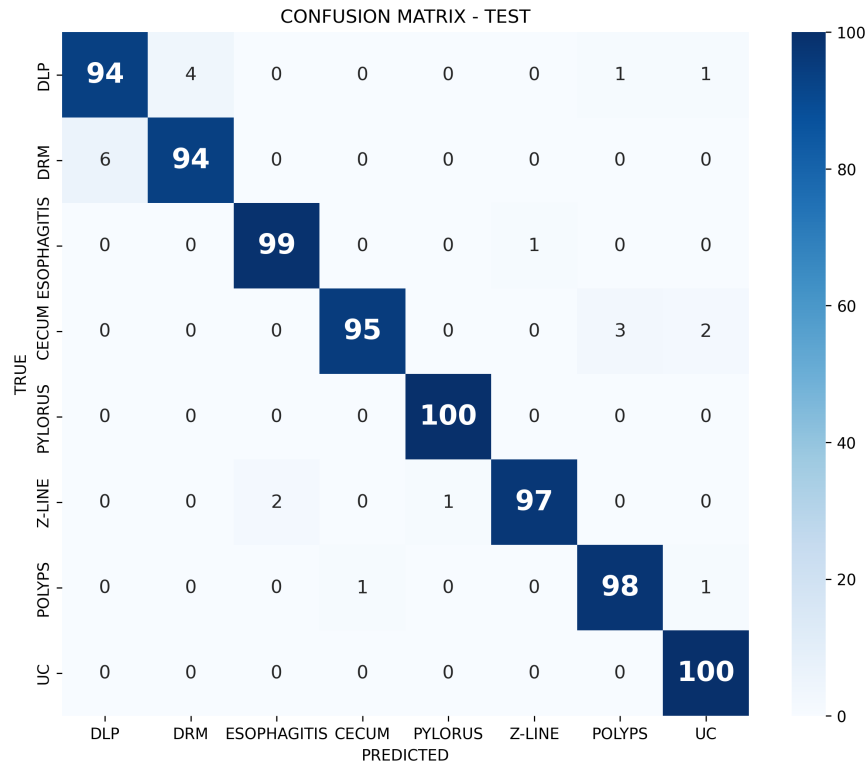


Figura 5: Matriz de confusão para o conjunto de dados de teste, destacando o desempenho da Resnet - 50 nas 8 classes: dyed-resection-margins, normal-cecum, normal-pylorus, linha Z normal, pólipos, ulcerative-colitis, esophagitis e dyed-lifted-polyps. Os elementos da diagonal representam previsões corretas, onde a classe prevista corresponde à classe verdadeira, indicando a precisão do modelo para cada categoria.

Para avaliar a generalização entre conjuntos de dados, aplicamos dois modelos treinados exclusivamente no Kvasir-v2: uma ResNet-50 regularizada, inicializada com pesos GastroNet, e uma ResNet-152 inicializada com pesos do ImageNet, ambos após o processo

de otimização e o treinamento final. Em seguida, avaliamos no CVC-ClinicDB [Bernal et al. \[2017\]](#), composto por 612 imagens com pólipos, onde a ResNet-50 atingiu acurácia de 0,9510 e a ResNet-152 atingiu 0,8513. Também avaliamos no Hyper-Kvasir-SEG [Borgli et al. \[2020\]](#), em um subconjunto de 1.000 imagens de pólipos, obtendo acurácia de 0,9780 para a ResNet-50 e 0,9600 para a ResNet-152.

5.3 Explicação por Blocos

Para facilitar a interpretação dos resultados, revisitamos a premissa central em relação às métricas de localização e deleção. O cenário ideal corresponde a blocos que apresentam pontuações baixas nas métricas de deleção (PF, RP, IROF), indicando um forte impacto causal na remoção de regiões marcadas como relevantes, combinado com valores altos nas métricas de localização (ALIR, PG, TKI), o que significa que a evidência utilizada pelo modelo está concentrada na lesão e não no contexto.

Vale ressaltar que, nesta etapa, conduzimos a análise de explicabilidade comparando apenas a ResNet-50 regularizada. A avaliação bloco a bloco em toda a ResNet-152 implicaria um custo computacional elevado, devido ao maior número de blocos e ao volume de curvas e métricas geradas por configuração.

5.3.1 Testes de Deleção

Nos testes de deleção, os blocos 13, 14 e 15 exibiram os menores valores médios para RP e IROF, indicando um refinamento progressivo nas regiões destacadas por esses blocos. Isso significa que a remoção dos pixels de maior relevância nessas etapas leva a curvas de deleção mais acentuadas e AUCs menores, validando assim a fidelidade causal de suas representações. De uma perspectiva global, entre os blocos 12 e 15, observamos uma redução no PF de 0,43 para 0,30 ($\Delta = -0.13$), no RP de 0,52 para 0,33 ($\Delta = -0.22$) e no IROF de 0,38 para 0,26 ($\Delta = -0.12$). As diferenças pareadas entre os blocos foram avaliadas usando o teste de Wilcoxon com correção de Holm ($p_{\text{Holm}} < 0.001$). Resultados detalhados estão disponíveis no Material Suplementar, Tabela 12.

Como análise complementar, a Figura 6 usa os blocos B11 e B12 como referência de comparação e mostra que, a partir do bloco B13, a probabilidade passa a cair mais cedo durante a deleção, indicando uma mudança no padrão das ativações nos blocos finais.

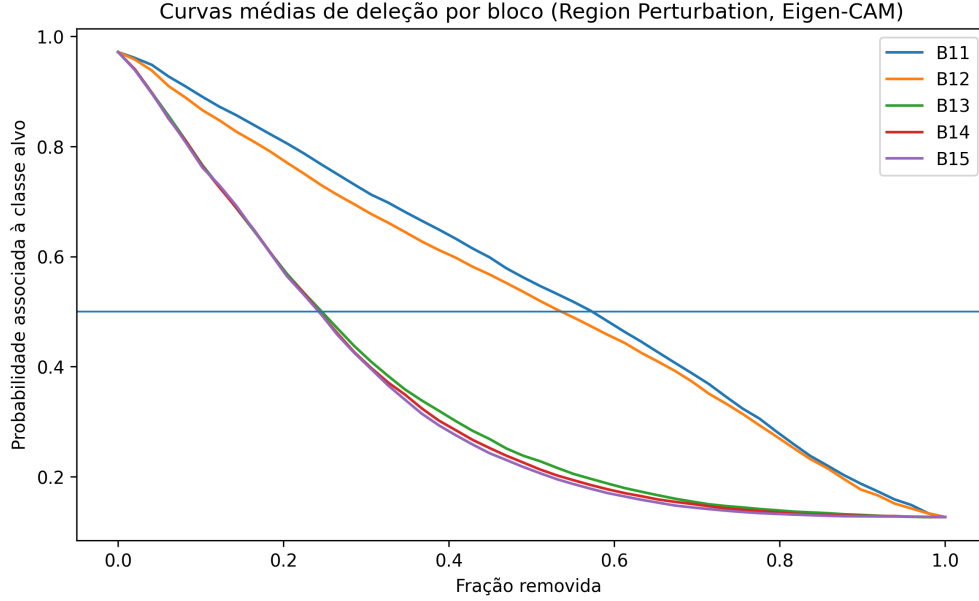


Figura 6: Curvas médias de deleção (Region Perturbation) para mapas Eigen-CAM por bloco, mostrando a evolução da probabilidade preditiva em função da fração removida. Os blocos finais (B13–B15) apresentam degradação mais precoce da probabilidade da classe alvo

Nos resultados estratificados por classe, o padrão permanece consistente com a análise global. Por exemplo, na classe de pólipos, entre os blocos 12 e 15, o PF diminui de 0,22 para 0,18 ($\Delta = -0.04$), o RP de 0,30 para 0,22 ($\Delta = -0.08$) e o IROF de 0,16 para 0,12 ($\Delta = -0.04$). Esse padrão persiste para as demais classes, com exceção da classe normal-z-line, na qual os blocos intermediários apresentaram melhor desempenho de deleção, com o PF atingindo seu mínimo no bloco 2 (0,09), o RP no bloco 6 (0,42) e o IROF no bloco 2 (0,21).

Em 17 das 24 combinações possíveis de classe e métrica, os menores valores de AUC ocorreram entre os blocos 13 e 15, corroborando o padrão observado. Portanto, os testes de deleção geralmente revelam, com exceção da classe normal-z-line, que os blocos 13, 14 e 15 concentram as ativações cuja remoção causa a degradação mais substancial no desempenho do modelo. Em outras palavras, tais ativações exibem maior fidelidade em relação ao mecanismo de tomada de decisão. Também é possível verificar que a AUC de deleção diminui consistentemente ao longo dos blocos finais, indicando um refinamento progressivo das características em direção à decisão do modelo.

5.3.2 Testes de Localização

Em termos de localização, analisamos três métricas complementares. PG é apresentado como uma proporção com Intervalos de Confiança (IC) de 95% de Wilson [Wilson \[1927\]](#); [McNemar \[1947\]](#), enquanto ALIR, ALOR e TKI empregam ICs de 95% de bootstrap percentil para a média. PG mede a fração de imagens em que o ponto de maior relevância no mapa de atribuição se encontra dentro da ROI. Portanto, valores altos indicam que o pico de atenção do modelo coincide com a lesão. A métrica ALIR quantifica a proporção de relevância positiva situada dentro da ROI, enquanto ALOR quantifica a proporção de relevância negativa fora da ROI. A métrica TKI mede a interseção entre os K pixels mais relevantes e a ROI, refletindo o quanto do suporte concentrado do modelo se sobrepõe às regiões anotadas.

De uma perspectiva global, os blocos 13, 14 e 15 emergiram como os de melhor desempenho na localização. Para as classes não difusas, observamos aumentos substanciais em todas as métricas ao comparar o bloco 12 com o bloco 15. A métrica PG melhorou de 0,47 para 0,78 ($\Delta = 0,32$; McNemar, $p_{\text{Holm}} < 0,001$, $d_z = -0,67$, $n = 397$). ALIR e TKI exibiram comportamentos semelhantes, mostrando aumentos de $\Delta = 0,27$ e $\Delta = 0,21$ respectivamente, com (Wilcoxon, $p_{\text{Holm}} < 0,001$).

Esse padrão indica que as evidências pró-classe migram progressivamente em direção à ROI ao longo dos blocos finais. A Tabela 8 detalha essas métricas para os blocos 12, 13, 14 e 15, incluindo os ICs de 95%.

Para polyps, dyed-lifted-polyps e dyed-resection-margins, as métricas PG, ALIR e TKI seguem a mesma tendência de aumento da concentração dentro da ROI, sugerindo, juntamente com os testes de deleção, que as regiões destacadas nos três últimos blocos são bem localizadas e causalmente importantes.

Para esophagitis, uma lesão com características difusas, observamos baixa localização e baixos efeitos de deleção, sugerindo ativações pró-classe mais fortes localizadas fora da ROI. A Figura 7 mostra que a concentração de relevância positiva dentro da ROI aumenta nos blocos finais para polyps e DRM, enquanto permanece baixa em esophagitis, reforçando o contraste entre classes com lesão bem delimitada e classes mais difusas.

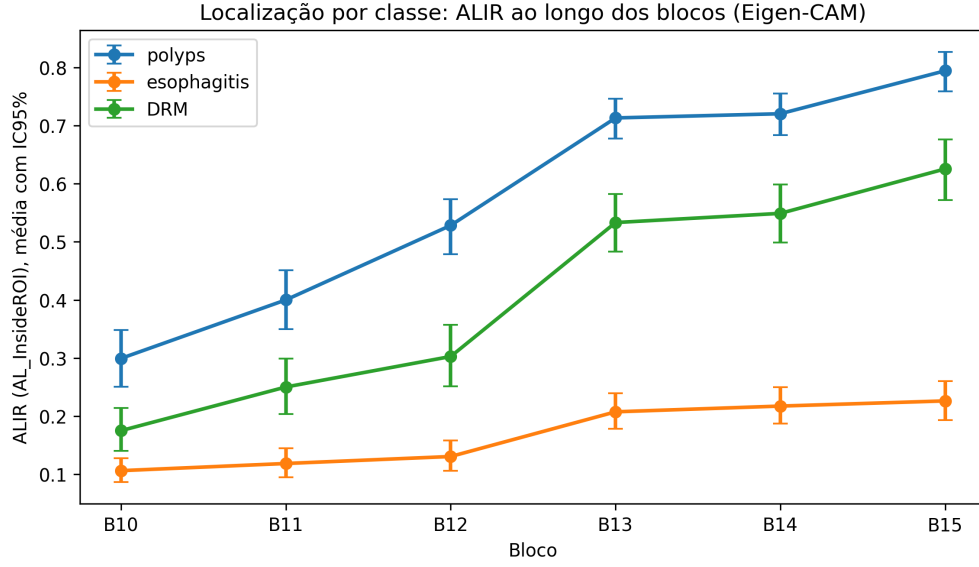


Figura 7: Localização por classe ao longo dos blocos finais: ALIR para Eigen-CAM, reportando a média e IC95% para polyps, esophagitis e DRM nos blocos B10–B15. Observa-se aumento progressivo de ALIR em polyps e DRM nos blocos finais, enquanto esophagitis mantém valores baixos, refletindo menor concentração de relevância positiva dentro da ROI.

Tabela 8: Localização por Classe e Bloco: AUC de Deleção (Médias) e Métricas de Localização [IC 95%].

CLASSE	BLOCO	RP	ALIR	PG	TKI
Polyps	12	0.30	[0.48–0.57]	[0.75–0.90]	[0.64–0.77]
Polyps	13	0.22	[0.68–0.75]	[0.96–1.00]	[0.78–0.86]
Polyps	14	0.22	[0.68–0.76]	[0.96–1.00]	[0.77–0.86]
Polyps	15	0.22	[0.76–0.83]	[0.94–1.00]	[0.77–0.85]
DLP	12	0.49	[0.35–0.44]	[0.40–0.60]	[0.40–0.57]
DLP	13	0.28	[0.65–0.73]	[0.89–0.98]	[0.79–0.88]
DLP	14	0.28	[0.68–0.76]	[0.89–0.98]	[0.79–0.88]
DLP	15	0.28	[0.75–0.83]	[0.89–0.98]	[0.79–0.88]
DRM	12	0.59	[0.25–0.36]	[0.36–0.56]	[0.31–0.47]
DRM	13	0.35	[0.48–0.58]	[0.82–0.94]	[0.58–0.71]
DRM	14	0.35	[0.50–0.60]	[0.80–0.93]	[0.58–0.70]
DRM	15	0.35	[0.57–0.68]	[0.80–0.93]	[0.58–0.70]
Esophagitis	12	0.44	[0.11–0.16]	[0.03–0.14]	[0.06–0.13]
Esophagitis	13	0.22	[0.18–0.24]	[0.18–0.34]	[0.17–0.26]
Esophagitis	14	0.21	[0.19–0.25]	[0.22–0.40]	[0.19–0.27]
Esophagitis	15	0.21	[0.19–0.26]	[0.24–0.42]	[0.19–0.27]

5.4 Comparação de Métodos de Explicabilidade

Na Tabela 9, observamos que os mapas LRP (R^-) alcançaram um ALOR agregado de 0,53. Em termos de interpretação, o ALOR quantifica a fração de relevância negativa concentrada no contexto extralesional, ou seja, em regiões que o modelo aprendeu a tratar como evidência de que essas áreas não correspondem à classe prevista.

Ao analisar os resultados por classe, observamos que esophagitis e dyed-resection-margins apresentaram pontuações de 0,79 e 0,65, respectivamente. Isso indica que, para essas classes, o modelo tende a se basear em evidências negativas fora da ROI para sua decisão. Em contraste, para dyed-lifted-polyps e polyps (0,36 e 0,34, respectivamente), observamos que o modelo utiliza menos relevância negativa fora da ROI para a predição.

Tabela 9: ALOR com LRP (R^-) Por classe [95% IC]

METRICA	METODO	CLASSE	VALOR
ALOR	LRP (R^-)	dyed-lifted-polyps	[0.31-0.40]
ALOR	LRP (R^-)	dyed-resection-margins	[0.59-0.69]
ALOR	LRP (R^-)	esophagitis	[0.75-0.82]
ALOR	LRP (R^-)	polyps	[0.30-0.38]
ALOR	LRP (R^-)	ALL	[0.50-0.56]

Para investigar o impacto da regularização nos métodos Grad-CAM e LRP, comparamos os checkpoints do ResNet-50 obtidos antes e depois do processo de regularização. O modelo regularizado apresentou 97,12% de acurácia, enquanto o modelo não regularizado alcançou 97,62% de acurácia.

Apesar dessa alta acurácia, os resultados apresentados na Tabela Suplementar 13 indicam que o modelo não regularizado apresentou um desempenho inferior em termos de localização. Onde o modelo não regularizado apresenta pior desempenho nas métricas de localização.

5.4.1 Efeito da Regularização

No modelo não regularizado, o LRP demonstrou maior fidelidade causal. Os testes de deleção confirmam essa vantagem, particularmente em Pixel Flipping ($\Delta = 0,268$, $d_z = 0,98$, $p < 0,001$), seguido por Region Perturbation e IROF (ambos com $d_z = 0,21$ e $p < 0,001$). No modelo regularizado, o Grad-CAM alcançou paridade estatística com o LRP nas métricas de deleção. Para Region Perturbation e IROF, as diferenças tornaram-se insignificantes

($\Delta \approx 0,003$, $p > 0,80$, indicando equivalência). A exceção foi Pixel Flipping, onde o LRP manteve uma vantagem, embora com magnitude reduzida em comparação com o cenário anterior ($\Delta = 0,089$, $d_z = 0,55$, $p < 0,001$).

A Tabela 10 detalha as métricas LRP (R^+) e Grad-CAM para o modelo regularizado. A Figura 8 evidencia que, no modelo não regularizado, Grad-CAM e LRP ocupam regiões distintas no espaço definido por ALIR e RP, enquanto, após a regularização, essa separação se reduz, indicando convergência no comportamento dos métodos.

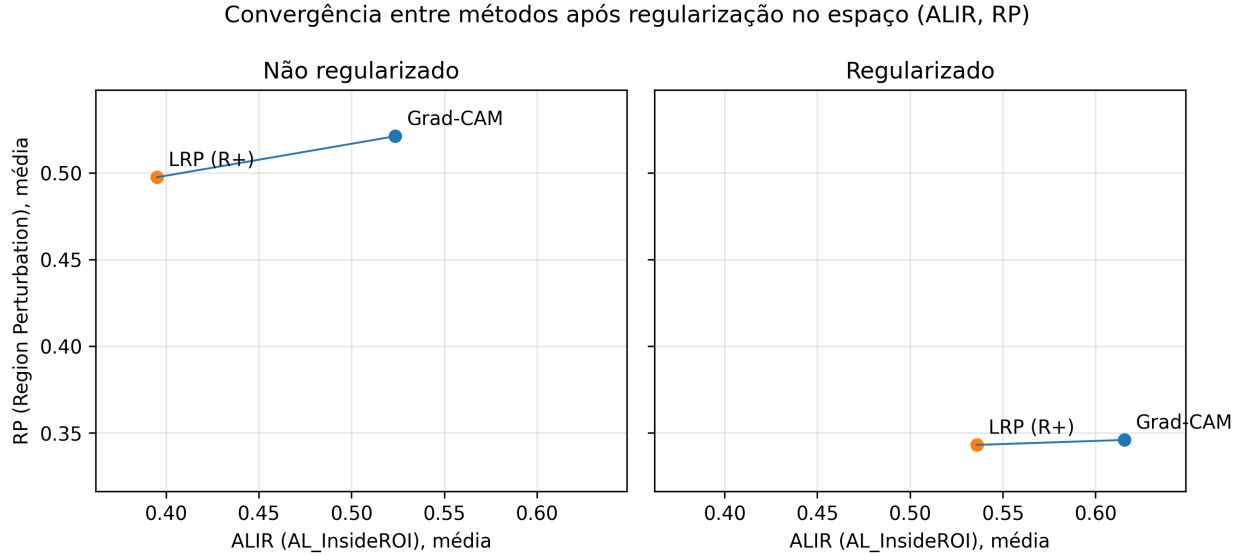


Figura 8: Convergência entre métodos de explicabilidade após regularização no espaço (ALIR, RP), considerando médias agregadas em todas as classes avaliadas. No modelo não regularizado, Grad-CAM e LRP ocupam regiões distintas, refletindo diferenças entre localização e impacto por deleção. Após a regularização, os pontos se aproximam, indicando redução da discrepância comportamental entre os métodos.

Nos resultados por classe com ROI usando o modelo regularizado, para polyps e dyed-lifted-polyps, o Grad-CAM reduz ligeiramente os escores de deleção em relação ao LRP (R^+), e as métricas de localização favorecem o Grad-CAM, que atinge valores de PG de até 0,95 em polyps e alcança um ALIR mais alto em todas as três classes. Em esophagitis, o LRP (R^+) atinge valores baixos em todas as três métricas de deleção, mas o PG e o ALIR permanecem baixos.

Para classes sem uma ROI delimitada, a comparação é restrita à deleção, onde o LRP (R^+) alcança AUCs de deleção mais baixos para essas classes, especialmente para normal-z-line e ulcerative-colitis. De modo geral, em ambos os modelos, o Grad-CAM produziu mapas

mais localizados sobre a lesão, enquanto o LRP (R^+) forneceu um sinal mais forte nos testes de deleção para o modelo não regularizado. Já no modelo regularizado, observou-se paridade estatística nos testes de deleção (resultados detalhados no Material Suplementar, 14 e 15).

As Figuras 9 e 10 demonstram que, após o processo de regularização, o modelo passou a concentrar a relevância de forma mais consistente dentro das regiões de interesse. Observa-se uma redução da ativação em áreas periféricas e artefatos visuais, com maior alinhamento entre os mapas de saliência e a ROI anotada. Esse comportamento indica menor dependência de pistas contextuais e maior uso de evidência visual associada diretamente à lesão, efeito corroborado pelo aumento das métricas de localização de relevância após a regularização.

Tabela 10: Resumo por classe e método: AUC de deleção [IC 95%] e localização (média).

CLASSE	METODO	PF	RP	IROF	PG
TODAS	GradCAM	[0.29–0.31]	[0.34–0.35]	[0.25–0.28]	0.78
TODAS	LRP (R^+)	[0.20–0.23]	[0.33–0.35]	[0.25–0.28]	0.68
Polyps	GradCAM	[0.16–0.18]	[0.21–0.23]	[0.11–0.12]	0.99
Polyps	LRP (R^+)	[0.06–0.07]	[0.23–0.26]	[0.13–0.15]	0.96
DLP	GradCAM	[0.16–0.18]	[0.27–0.30]	[0.11–0.15]	0.95
DLP	LRP (R^+)	[0.12–0.15]	[0.27–0.30]	[0.14–0.18]	0.80
DRM	GradCAM	[0.48–0.54]	[0.36–0.42]	[0.62–0.69]	0.87
DRM	LRP (R^+)	[0.52–0.58]	[0.32–0.37]	[0.63–0.70]	0.81
Esophagitis	GradCAM	[0.20–0.23]	[0.20–0.23]	[0.11–0.13]	0.30
Esophagitis	LRP (R^+)	[0.04–0.05]	[0.19–0.22]	[0.11–0.13]	0.14
Cecum	GradCAM	[0.12–0.13]	[0.28–0.30]	[0.12–0.13]	–
Cecum	LRP (R^+)	[0.18–0.20]	[0.32–0.35]	[0.11–0.13]	–
Pylorus	GradCAM	[0.31–0.34]	[0.39–0.41]	[0.26–0.29]	–
Pylorus	LRP (R^+)	[0.10–0.11]	[0.41–0.44]	[0.27–0.31]	–
z-line	GradCAM	[0.42–0.46]	[0.47–0.52]	[0.37–0.41]	–
z-line	LRP (R^+)	[0.09–0.10]	[0.42–0.45]	[0.26–0.30]	–
UC	GradCAM	[0.44–0.48]	[0.45–0.48]	[0.31–0.35]	–
UC	LRP (R^+)	[0.51–0.55]	[0.44–0.47]	[0.31–0.35]	–

5.4.2 Comparação com a literatura (ALIR/PG)

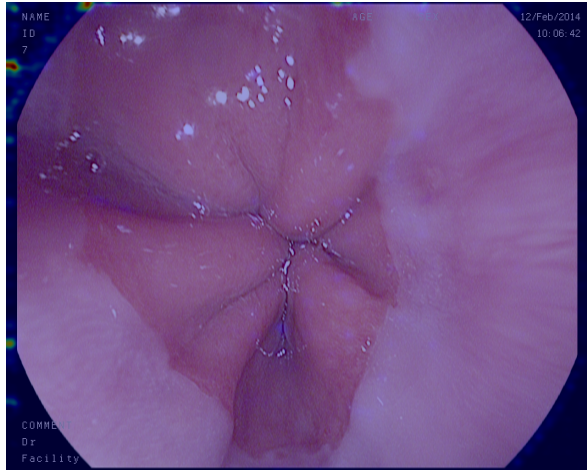
ALIR mede a fração de relevância positiva contida na ROI, assim, um aumento de $\Delta = 0,1$ representa +10 pontos percentuais de evidência pró-classe concentrada na lesão, reduzindo dependência do contexto. Embora não existam baselines consolidados em endoscopia, trabalhos em outras modalidades que empregam RMA/ALIR reportam valores na

ordem de 0,74 em seus melhores resultados, enquanto Pointing Game (PG) tipicamente varia conforme modalidade e tarefa.

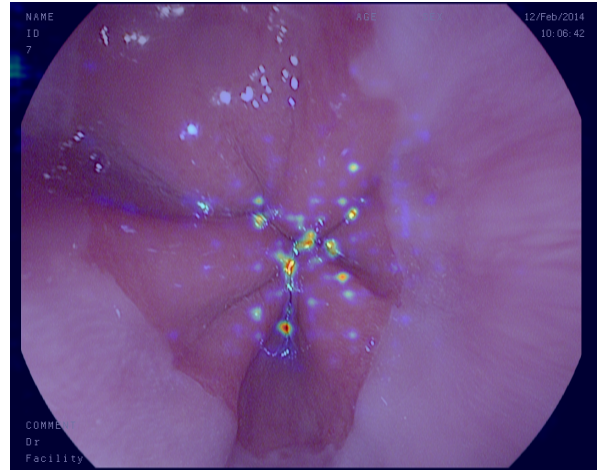
Na Tabela 11, o método proposto atinge ALIR de 0,74 com Grad-CAM e 0,65 com LRP, além de PG de até 0,99 em pólipos, posicionando-se de forma competitiva em relação à literatura. Finalmente, a interpretação é reforçada pela concordância entre localização e deleção nos blocos finais, onde ALIR aumenta ao mesmo tempo em que as AUCs de deleção diminuem, evidenciando alinhamento espacial e impacto causal da evidência selecionada. Por fim, ressaltamos que a comparação com estudos de outras modalidades tem caráter de referência de alto nível, útil para contextualizar magnitudes, mas não permite comparações estritas devido a diferenças de protocolo, anotação de ROI e distribuição de dados.

Tabela 11: Comparação de trabalhos por métrica, método de explicabilidade e modalidade.

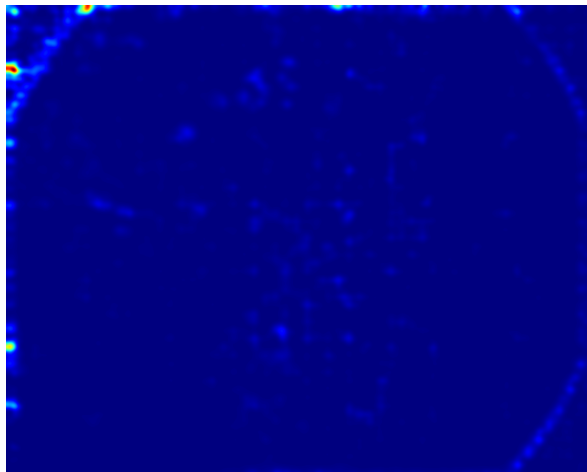
Trabalho	Métrica	Método	Resultado	Deleção	Modalidade	Classe
Saporta et al. [2022]	PG	Grad-CAM	0.91	Não	CXR	Cardiomegalia
Al-Hammouri et al. [2026]	PG	Grad-CAM	0.91	Não	CT	Hemorragia
He et al. [2025]	PG	Grad-CAM	0.81	Não	Múltiplas	Múltiplas
Srinivasan et al. [2022]	ALIR	LRP	0.74	Não	CFP	Hemorragia
Pahde et al. [2025]	ALIR	LRP	0.74	Não	Artefatos	data/hora
Siegel et al. [2025]	ALIR	SmoothGrad	0.31	Não	MRI	binária
Método Proposto	PG	Grad-CAM	0.99	Sim	Endoscopia	Pólipos
Método Proposto	ALIR	Grad-CAM	0.74	Sim	Endoscopia	Pólipos
Método Proposto	PG	LRP	0.96	Sim	Endoscopia	Pólipos
Método Proposto	ALIR	LRP	0.65	Sim	Endoscopia	Pólipos



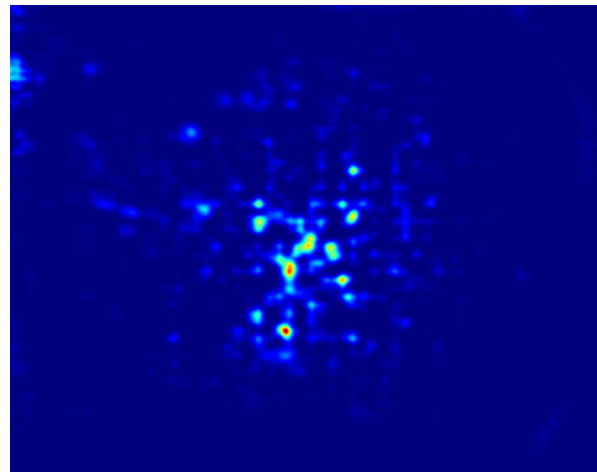
(a) Não regularizado: overlay



(b) Regularizado: overlay

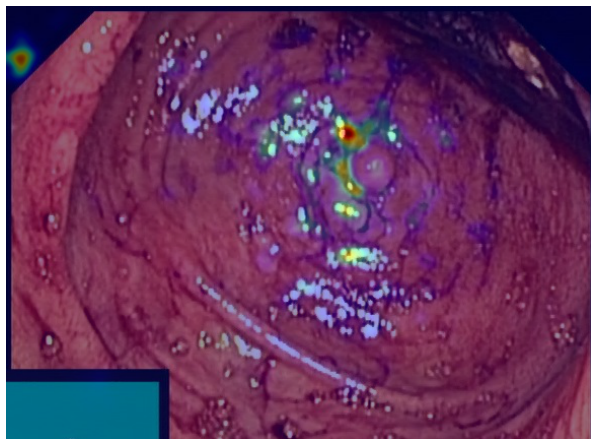


(c) Não regularizado: heatmap

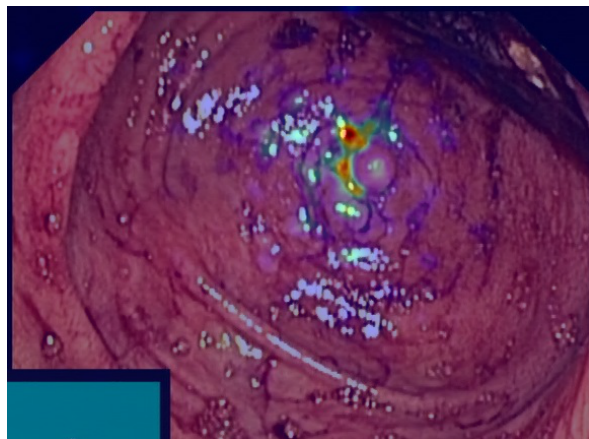


(d) Regularizado: heatmap

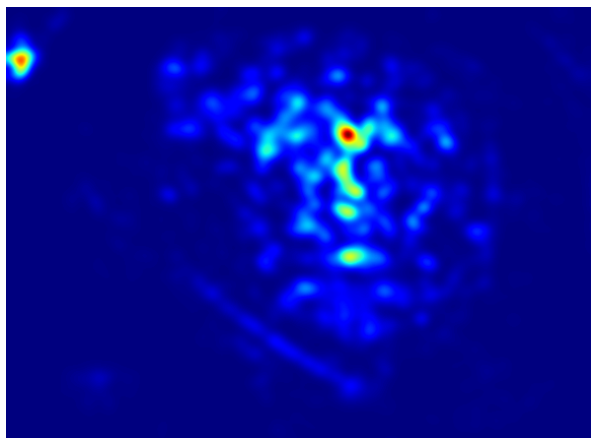
Figura 9: Comparação visual dos mapas LRP para esophagitis. Coluna esquerda: modelo não regularizado. Coluna direita: modelo regularizado. Primeira linha: overlay na imagem original. Segunda linha: heatmap isolado.



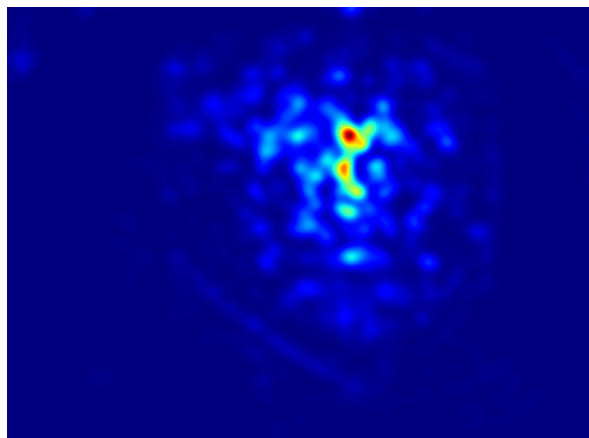
(a) Não regularizado: overlay



(b) Regularizado: overlay



(c) Não regularizado: heatmap



(d) Regularizado: heatmap

Figura 10: Comparação visual dos mapas LRP para pólipos. Coluna esquerda: modelo não regularizado. Coluna direita: modelo regularizado. Primeira linha: overlay na imagem original. Segunda linha: heatmap isolado.

6 CAPÍTULO 6: CONCLUSÕES E TRABALHOS FUTUROS

6.1 DISCUSSÃO

Os resultados obtidos destacam a efetividade do método de otimização proposto. A acurácia alcançada no conjunto de teste (97,12%) representa um desempenho próximo aos melhores resultados encontrados na literatura utilizando todas as classes do conjunto de dados Kvasir v2, sem a utilização de múltiplos modelos para predição, como em [Mary et al. \[2023\]](#), que obteve 98,02%, com uma diferença de apenas 0,9%.

A consistência entre os resultados de treinamento e teste demonstra a ausência de overfitting, evidenciando a capacidade de generalização do modelo. Os resultados sugerem que a inicialização com pesos mais específicos do domínio pode favorecer a generalização fora do domínio, especialmente quando combinada com o nosso processo de otimização. No CVC-ClinicDB, a ResNet-50 inicializada com pesos GastroNet ficou 0,0997 acima em acurácia da ResNet-152 inicializada com ImageNet, o que é compatível com a hipótese de que representações pré-treinadas em endoscopia capturam texturas, iluminação e padrões mucosos mais próximos dos alvos clínicos.

No Hyper-Kvasir-SEG a diferença foi menor, levando em conta que embora Kvasir-v2 e Hyper-Kvasir-SEG sejam conjuntos distintos, ambos foram coletados no mesmo hospital, o que indica que o ganho tende a crescer quando a mudança de domínio é mais forte, como entre Kvasir-v2 e CVC-ClinicDB. Os resultados de validação externa são compatíveis com os achados de [Boers et al. \[2024\]](#), que observaram ganho médio de 1,63% ao usar pré-treinamento auto-supervisionado no domínio, superando o pré-treinamento supervisionado iniciados com imagens naturais para a ResNet-50. No cenário específico deste trabalho, a diferença tende a ficar mais evidente porque os modelos não foram ajustados nas bases avaliadas, onde a inferência foi realizada em regime de transferência direta, sem fine-tuning adicional no CVC-ClinicDB ou Hyper-Kvasir-SEG.

No caso avaliado em nosso trabalho, a ResNet-50 inicializada com pesos do GastroNet e refinada pelo nosso processo de otimização superou a ResNet-152 inicializada com ImageNet no CVC-ClinicDB em 0,0997 de acurácia, e apresentou diferença menor no Hyper-Kvasir-SEG. Reforçando a hipótese de que pesos mais específicos do domínio favorecem a generalização sob mudanças moderadas de domínio, aqui caracterizadas por dados do mesmo tipo de exame, porém coletados em ambientes distintos, com variações de equipamento, iluminação, protocolo e população. No tocante a explicabilidade os resultados indicam que no modelo não regularizado, houve dissociação clara entre localização e fidelidade causal, onde o Grad-CAM demonstrou concentrar evidência na ROI, enquanto o LRP (R^+) apresentou

vantagem nos testes de deleção. Isso sugere que mapas visualmente plausíveis podem não refletir com a mesma intensidade, a dependência causal do preditor. Esse achado é compatível com o estudo de [Adebayo et al. \[2018\]](#)

Após a regularização, constataou-se a paridade estatística entre Grad-CAM e LRP nas deleções, exceto em Pixel Flipping, sugerindo um efeito de alinhamento entre o comportamento do modelo e a sensibilidade capturada por gradientes. A nossa hipótese é que a regularização reduz o espaço de soluções que dependem de pistas dispersas, facilitando que métodos baseados em gradiente recuperem sinais mais próximos do mecanismo decisório. Ao custo de 0,50 p.p. de acurácia com a regularização, observou-se ganho de consistência explicativa, o que é relevante quando o objetivo inclui auditoria e não apenas desempenho.

O padrão moderado de relevância negativa extralesional, com ALOR agregado de 0,53 no LRP (R^-), adiciona uma camada comportamental que a acurácia não captura. Em classes como esophagitis (0,79) e dyed-resection-margins (0,65), a decisão indica incorporar contraste fora da ROI, o que pode refletir um mecanismo legítimo de negação espacial, ou dependência contextual do tipo “Clever Hans”.

6.2 CONCLUSÕES

Este estudo demonstrou que a otimização em blocos do ResNet-50 via algoritmo genético permite um desempenho competitivo na classificação endoscópica, mantendo a generalização fora do domínio de treinamento. Especificamente, além de atingir 97,12% de acurácia no conjunto de teste Kvasir-v2, os modelos treinados no Kvasir-v2 mantiveram alto desempenho quando aplicado a conjuntos de dados externos, atingindo 95,10% de acurácia no CVC-ClinicDB e 97,80% no subconjunto de pólipos do Hyper-Kvasir-SEG. Em relação à explicabilidade, os resultados indicaram que os blocos finais, particularmente os blocos 13, 14 e 15, concentram as ativações mais relevantes para a decisão.

A comparação entre os métodos apontou para papéis complementares: o Grad-CAM apresentou maior alinhamento médio com a ROI, favorecendo a auditoria visual, enquanto o LRP (R^+) apresentou maior fidelidade causal no modelo não regularizado, uma vantagem que se dissipou estatisticamente com a regularização. Complementarmente, o LRP (R^-) evidenciou um padrão moderado de relevância negativa no contexto extralesional.

Em trabalhos futuros, será necessário investigar se os blocos que ativam fora da ROI atuam como um mecanismo de negação, delimitando a decisão por meio de contraste espacial, ou se refletem a dependência contextual através de atalhos do tipo “Clever Hans”. Nesse sentido, um diferencial deste trabalho é que a abertura da “caixa preta” nos permitiu

delimitar o escopo onde a Resnet-50 e os métodos avaliados são mais confiáveis, e onde expõem vulnerabilidades. Particularmente, nossa análise revelou um padrão comportamental confiável para classes não difusas, marcado por foco predominante na ROI e alta fidelidade causal, ao mesmo tempo que destacou a necessidade de uma investigação mais aprofundada em classes difusas. Essas são descobertas que métricas de acurácia isoladas tendem a obscurecer.

Além disso, recomendamos as seguintes direções para trabalhos futuros: propor e avaliar métodos para investigar o comportamento do modelo em classes de ativação difusa, como classes normais, ulcerative-colitis e esophagitis, bem como avaliar arquiteturas alternativas com a remoção seletiva de blocos redundantes.

REFERÊNCIAS

- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., e Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., e Kim, B. (2018). Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, página 9525–9536, Red Hook, NY, USA. Curran Associates Inc.
- Ahmed, I. A., Senan, E. M., e Shatnawi, H. S. A. (2023). Hybrid models for endoscopy image analysis for early detection of gastrointestinal diseases based on fused features. *Diagnostics*, 13(10):1758.
- Akinrinmade, A. O., Adebile, T. M., Ezuma-Ebong, C., Bolaji, K., Ajufo, A., Adigun, A. O., Mohammad, M., Dike, J. C., e Okobi, O. E. (2023). Artificial intelligence in healthcare: Perception and reality. *Cureus*, 15(9):e45594.
- Al-Hammouri, M. F., Vamsi, B., Almalkawi, I. T., e Al Bataineh, A. (2026). X-hem: An explainable and trustworthy ai-based framework for intelligent healthcare diagnostics. *Computers*, 15(1).
- Al-Otaibi, S., Rehman, A., Mujahid, M., Alotaibi, S., e Saba, T. (2024). Efficient-gastro: Optimized efficientnet model for the detection of gastrointestinal disorders using transfer learning and wireless capsule endoscopy images. *PeerJ Computer Science*, 10:e1902.
- Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y. B., Matuszewski, B., Gridach, M., Voiculescu, I., Yoganand, V., Chavan, A., Raj, A., Nguyen, N. T., Tran, D. Q., Huynh, L. D., Boutry, N., Rezdy, S., Chen, H., Choi, Y. H., Subramanian, A., Balasubramanian, V., Gao, X. W., Hu, H., Liao, Y., Stoyanov, D., Daul, C., Realdon, S., Cannizzaro, R., Lamarque, D., Tran-Nguyen, T., Bailey, A., Braden, B., East, J. E., e Rittscher, J. (2021). Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis*, 70:102002.
- Alshardan, A., Saeed, M. K., Alotaibi, S. D., Alashjaee, A. M., Salih, N., e Marzouk, R. (2024). Harbor seal whiskers optimization algorithm with deep learning-based medical imaging analysis for gastrointestinal cancer detection. *Health Information Science and Systems*, 12(1):35.

- Anders, C. J., Neumann, D., Samek, W., Müller, K., e Lapuschkin, S. (2021). Software for dataset-wide xai: From local explanations to global insights with zennit, corelay, and virelay.
- Auzine, M. M., Heenaye-Mamode Khan, M., Baichoo, S., Gooda Sahib, N., Bissoonauth-Daiboo, P., Gao, X., e Heetun, Z. (2024). Development of an ensemble cnn model with explainable ai for the classification of gastrointestinal cancer. *PLoS One*, 19(6):e0305628.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., e Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.
- Bernal, J., Tajkbaksh, N., Sánchez, F. J., Matuszewski, B. J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S. R., Fernández-Esparrach, G., Dray, X., Liang, J., e Histace, A. (2017). Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, 36(6):1231–1249.
- Boers, T. G. W., Fockens, K. N., van der Putten, J. A., Jaspers, T. J. M., Kusters, C. H. J., Jukema, J. B., Jong, M. R., Struyvenberg, M. R., de Groof, J., Bergman, J. J., de With, P. H. N., e van der Sommen, F. (2024). Foundation models in gastrointestinal endoscopic ai: Impact of architecture, pre-training approach and data efficiency. *Medical Image Analysis*, 98:103298.
- Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., e de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283.
- Elsken, T., Metzen, J. H., e Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21.
- Fati, S. M., Senan, E. M., e Azar, A. T. (2022). Hybrid and deep learning approach for early diagnosis of lower gastrointestinal diseases. *Sensors*, 22(11):4079.

- Fong, R. C. e Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, páginas 3449–3457.
- Fortin, F.-A., De Rainville, D., Gardner, M.-A., Parizeau, M., e Gagné, C. (2012). Deap: Evolutionary algorithms made easy. In *Journal of Machine Learning Research*, volume 13, páginas 2171–2175.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., e Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., e Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42.
- Gunasekaran, H., Ramalakshmi, K., Swaminathan, D. K., J, A., e Mazzara, M. (2023). Git-net: An ensemble deep learning-based gi tract classification of endoscopic images. *Bioengineering*, 10(7):809.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.
- He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 770–778.
- He, L., Luan, L., e Hu, D. (2025). Deep learning-based image classification for ai-assisted integration of pathology and radiology in medical imaging. *Frontiers in Medicine*, Volume 12 - 2025.
- Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., Lapuschkin, S., e Höhne, M. M.-C. (2023). Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *J. Mach. Learn. Res.*, 24(1).

- Hossain, T., Shamrat, F. M. J. M., Zhou, X., Mahmud, I., Mazumder, M. S. A., Sharmin, S., e Gururajan, R. (2024). Development of a multi-fusion convolutional neural network (mf-cnn) for enhanced gastrointestinal disease diagnosis in endoscopy image analysis. *PeerJ Computer Science*, 10:e1950.
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.
- Jong, M. R., Boers, T. G. W., Fockens, K. N., Jukema, J. B., Kusters, C. H. J., Jaspers, T. J. M., van Eijck van Heslinga, R. A. H., Slooter, F. C., Struyvenberg, M. R., Bisschops, R., van der Putten, J. A., de With, P. H. N., van der Sommen, F., de Groof, A. J., e Bergman, J. J. G. H. M. (2025). GastroNet-5M: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. *Gastroenterology*. Advance online publication.
- KatherLab e Kather, J. N. (2022). Impact of text overlay in endoscopic images on bias in deep learning classification. *Scientific Reports*.
- Khan, M. A., Sahar, N., Khan, W. Z., Alhaisoni, M., Tariq, U., Zayyan, M. H., Kim, Y. J., e Chang, B. (2022). Gestronet: A framework of saliency estimation and optimal deep learning features based gastrointestinal diseases detection and classification. *Diagnostics*, 12(11):2718.
- Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., e Lapuschkin, S. (2020). Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, páginas 1–7.
- Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, páginas 1097–1105.
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., e Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.
- LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Letafatkar, N., El-Sehrawy, A. A. M. A., Prasad, K., Alkhayyat, A., Amini-Salehi, E., Hasanpour, M., Taleshani, M. N., Hashemi, M., Alotaibi, H., Rashidian, P., Keivanlou, M.-H., e Hassanipour, S. (2025). Artificial intelligence in endoscopy and colonoscopy: a comprehensive bibliometric analysis of global research trends. *Frontiers in Medicine*, Volume 12 - 2025.

- Lipton, Z. C. (2017). The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., e Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- maintainers, T. e contributors (2016). TorchVision: PyTorch’s Computer Vision library.
- Mary, A. A., Raj, A. P. W., Karthik, C., e Karunaharan, A. (2023). Gastronet: A custom deep learning approach for classification of anomalies in gastrointestinal endoscopy images. *Current Medical Imaging*. Advance online publication.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., e Müller, K.-R. (2019). *Layer-Wise Relevance Propagation: An Overview*, páginas 193–209. Springer International Publishing, Cham.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., e Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.
- Muhammad, M. B. e Yeasin, M. (2020). Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, páginas 1–7.
- Mukhtorov, D., Rakhmonova, M., Muksimova, S., e Cho, Y. I. (2023). Endoscopic image classification based on explainable deep learning. *Sensors*, 23(6):3176.
- Nadimi, E. S., Braun, J.-M., Schelde-Olesen, B., Khare, S., Gogineni, V. C., Blanes-Vidal, V., e Baatrup, G. (2025). Towards full integration of explainable artificial intelligence in colon capsule endoscopy’s pathway. *Scientific Reports*, 15(1):5960.
- Nakashima, J. e Duong, H. (2025). *Radiology, Image Production and Evaluation*. StatPearls Publishing, Treasure Island (FL). Updated 2023 Jul 31. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK553145/>.
- Nam, W., Gur, S., Choi, J., Wolf, L., e Lee, S.-W. (2020). Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2501–2508.

- Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836.
- Pahde, F., Wiegand, T., Lapuschkin, S., e Samek, W. (2025). Ensuring medical ai safety: interpretability-driven detection and mitigation of spurious model behavior and associated data. *Machine Learning*, 114(9):206.
- Pan, S. J. e Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
- Petsiuk, V., Das, A., e Saenko, K. (2018). RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, página 151. BMVA Press.
- Pincirolì Vago, N. O., Milani, F., Fraternali, P., e da Silva Torres, R. (2021). Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. *Journal of Imaging*, 7(7).
- Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., e Halvorsen, P. (2017). Kvasir: A multi-class image dataset for computer-aided gastrointestinal disease detection. In *Proceedings of the 8th ACM Multimedia Systems Conference*, páginas 164–169. ACM.
- Qin, K., Li, J., Fang, Y., Xu, Y., Wu, J., Zhang, H., Li, H., Liu, S., e Li, Q. (2022). Convolution neural network for the diagnosis of wireless capsule endoscopy: A systematic review and meta-analysis. *Surgical Endoscopy*, 36(1):16–31.
- Real, E., Aggarwal, A., Huang, Y., e Le, Q. V. (2019). Regularized evolution for image classifier architecture search. arXiv preprint arXiv:1802.01548.
- Redmon, J. e Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- Ribeiro, M. T., Singh, S., e Guestrin, C. (2016). "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 1135–1144.

- Rieger, L. e Hansen, L. (2020). Irof: a low resource evaluation metric for explanation methods. In *Proceedings of the Workshop AI for Affordable Healthcare at ICLR 2020*. Workshop AI for Affordable Healthcare at ICLR 2020 ; Conference date: 26-04-2020 Through 26-04-2020.
- Rieger, L., Singh, C., Murdoch, W., e Yu, B. (2020). Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, páginas 8116–8126. International Machine Learning Society (IMLS). 37th International Conference on Machine Learning, ICML 2020 ; Conference date: 13-07-2020 Through 18-07-2020.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., e Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.
- Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P., e Rajpurkar, P. (2022). Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., e Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, páginas 618–626.
- Sharma, A., Kumar, R., e Garg, P. (2023). Deep learning-based prediction model for diagnosing gastrointestinal diseases using endoscopy images. *International Journal of Medical Informatics*, 177:105142.
- Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., e Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.
- Shorten, C. e Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

- Siegel, N. T., Cole, J. H., Habes, M., Haufe, S., Ritter, K., e Schulz, M.-A. (2025). Explainable ai methods for neuroimaging: Systematic failures of common tools, the need for domain-specific validation, and a proposal for safe application.
- Srinivasan, V., Strodthoff, N., Ma, J., Binder, A., Müller, K.-R., e Samek, W. (2022). To pretrain or not? a systematic analysis of the benefits of pretraining in diabetic retinopathy. *PLOS ONE*, 17(10):1–18.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., e Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.
- Sutton, R. T., Zaïane, O. R., Goebel, R., e Baumgart, D. C. (2022). Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Scientific Reports*, 12(1):2748.
- Terven, J. e Cordova-Esparza, D. (2023). A comprehensive review of yolo: From yolov1 and beyond. arXiv preprint arXiv:2304.00501.
- Theiner, J., Müller-Budack, E., e Ewerth, R. (2022). Interpretable semantic photo geo-location. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, páginas 1474–1484.
- Vakanski, A., Xian, M., e Freer, P. (2020). Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound in Medicine Biology*, 46.
- Varam, D., Mitra, R., Mkadmi, M., Riyas, R. A., Abuhani, D. A., Dhou, S., e Alzaatreh, A. (2023). Wireless capsule endoscopy image classification: An explainable ai approach. *IEEE Access*, 11:105262–105280.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.
- World Medical Association (2024). Wma declaration of helsinki – ethical principles for medical research involving human participants. <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-s> Adopted by the 18th WMA General Assembly, Helsinki, 1964, and amended by the 75th WMA General Assembly, Helsinki, 2024. Accessed on August 2, 2025.
- Yosinski, J., Clune, J., Bengio, Y., e Lipson, H. (2014). How transferable are features in deep neural networks? *arXiv eprint:1411.1792*.

- Zennit Contributors (2025). Zennit documentation. <https://zennit.readthedocs.io/>. Accessed: Oct. 22, 2025.
- Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., e Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126:1084–1102.
- Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Naseem, U., e Gu, Y. (2025). Robust and explainable framework to address data scarcity in diagnostic imaging. *Computers in Biology and Medicine*, 197:111052.
- Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., e Torralba, A. (2015). Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 2921–2929.

ANEXO A – ANEXOS E APÊNDICES 1

A Tabela 12 quantifica, no modelo regularizado, a diferença pareada entre as atribuições do EigenCAM extraídas do bloco final (B15) e de um bloco intermediário (B12). Para cada classe e métrica, reportamos Δ como a diferença média pareada (B15 menos B12), o p_{Holm} corrigido por Holm, que considera múltiplas comparações, o tamanho de efeito pareado d_z e o número de pares válidos n . Valores negativos de Δ para métricas de deleção (PF, RP, IROF) indicam menor AUC de deleção em B15, consistente com maior fidelidade causal das regiões destacadas no bloco final. Valores positivos de Δ para métricas de localização (ALIR, PG, TKI) indicam maior alinhamento espacial com a ROI em B15.

A Tabela 13 detalha, para o modelo não regularizado, a localização por classe e por bloco: AUC de deleção (médias) e métricas de localização [IC 95%].

Tabela 12: Comparação de blocos no modelo regularizado: B15 - B12 (Eigen-CAM). Reportado como Δ (diferença média), p_{Holm} , d_z e n .

CLASSE	METRICA	Δ	p_{Holm}	d_z	n
TODAS	PF	-0.129	$< 10^{-3}$	0.64	800
TODAS	RP	-0.192	$< 10^{-3}$	0.88	800
TODAS	IROF	-0.119	$< 10^{-3}$	0.79	800
TODAS	ALIR	0.270	$< 10^{-3}$	-1.32	397
TODAS	PG	0.317	$< 10^{-3}$	-0.67	397
TODAS	TKI	0.213	$< 10^{-3}$	-0.67	397
Polyps	PF	-0.043	0.059	0.22	100
Polyps	RP	-0.077	1.000	0.38	100
Polyps	IROF	-0.037	1.000	0.28	100
Polyps	ALIR	0.266	$< 10^{-3}$	-1.52	99
Polyps	PG	0.152	0.003	-0.42	99
Polyps	TKI	0.103	0.649	-0.38	99
DLP	PF	-0.181	0.002	0.70	100
DLP	RP	-0.212	$< 10^{-3}$	0.85	100
DLP	IROF	-0.108	0.006	0.63	100
DLP	ALIR	0.394	$< 10^{-3}$	-1.95	100
DLP	PG	0.450	$< 10^{-3}$	-0.90	100
DLP	TKI	0.358	$< 10^{-3}$	-0.90	100
DRM	PF	-0.220	$< 10^{-3}$	1.22	100
DRM	RP	-0.235	$< 10^{-3}$	0.87	100
DRM	IROF	-0.171	$< 10^{-3}$	1.31	100
DRM	ALIR	0.323	$< 10^{-3}$	-1.77	98
DRM	PG	0.418	$< 10^{-3}$	-0.84	98
DRM	TKI	0.255	$< 10^{-3}$	-0.79	98
Esophagitis	PF	-0.094	$< 10^{-3}$	0.62	100
Esophagitis	RP	-0.227	$< 10^{-3}$	1.19	100
Esophagitis	IROF	-0.124	$< 10^{-3}$	1.02	100
Esophagitis	ALIR	0.096	$< 10^{-3}$	-0.80	100
Esophagitis	PG	0.250	$< 10^{-3}$	-0.55	100
Esophagitis	TKI	0.137	$< 10^{-3}$	-0.79	100

Tabela 13: Localização no modelo não regularizado: AUC de deleção (médias) e métricas de localização [IC 95%].

CLASSE	BLOCK	RP	ALIR	PG	TKI
Esophagitis	13	0.38	[0.14–0.19]	[0.07–0.20]	[0.11–0.17]
Esophagitis	14	0.37	[0.13–0.18]	[0.06–0.17]	[0.11–0.18]
Esophagitis	15	0.37	[0.13–0.18]	[0.08–0.21]	[0.11–0.18]
Polyps	13	0.60	[0.31–0.40]	[0.59–0.77]	[0.50–0.65]
Polyps	14	0.51	[0.45–0.53]	[0.79–0.92]	[0.65–0.77]
Polyps	15	0.49	[0.56–0.63]	[0.85–0.96]	[0.70–0.79]
DLP	13	0.59	[0.36–0.44]	[0.56–0.75]	[0.48–0.64]
DLP	14	0.52	[0.47–0.55]	[0.70–0.86]	[0.63–0.76]
DLP	15	0.49	[0.57–0.66]	[0.81–0.94]	[0.72–0.83]
DRM	13	0.46	[0.28–0.38]	[0.67–0.84]	[0.50–0.65]
DRM	14	0.39	[0.40–0.50]	[0.82–0.94]	[0.59–0.72]
DRM	15	0.38	[0.47–0.57]	[0.82–0.94]	[0.58–0.71]

A Tabela 14 apresenta, para o modelo regularizado, estatísticas descritivas por classe e por método, cobrindo métricas de deleção (PF, RP, IROF) e métricas de localização (ALIR, PG, TKI). Para cada combinação, reportamos a média juntamente com o intervalo de confiança de 95% (IC 95%). Para classes sem anotações de ROI, as métricas de localização são reportadas como indisponíveis, pois exigem explicitamente máscaras de ROI. Esta tabela dá suporte à caracterização por método e por classe, enquanto as comparações inferenciais pareadas entre métodos são reportadas na Tabela 15.

Tabela 14: Resumo por classe e método (modelo regularizado): média [IC 95%] para métricas de deleção e de localização.

CLASSE	METHOD	PF	RP	IROF	ALIR	PG	TKI
TODAS	GradCAM	0.302 [0.29–0.31]	0.346 [0.34–0.35]	0.267 [0.25–0.28]	0.615 [0.58–0.65]	0.776 [0.73–0.81]	0.608 [0.58–0.64]
TODAS	LRP (R ⁺)	0.213 [0.20–0.23]	0.343 [0.33–0.35]	0.262 [0.25–0.28]	0.536 [0.51–0.56]	0.675 [0.63–0.72]	0.474 [0.45–0.50]
Polyps	GradCAM	0.169 [0.16–0.18]	0.221 [0.21–0.23]	0.116 [0.11–0.12]	0.741 [0.70–0.78]	0.990 [0.94–1.00]	0.778 [0.73–0.82]
Polyps	LRP (R ⁺)	0.063 [0.06–0.07]	0.247 [0.23–0.26]	0.139 [0.13–0.15]	0.656 [0.63–0.68]	0.960 [0.90–0.98]	0.515 [0.48–0.55]
DLP	GradCAM	0.170 [0.16–0.18]	0.286 [0.27–0.30]	0.130 [0.11–0.15]	0.808 [0.77–0.85]	0.950 [0.89–0.98]	0.821 [0.77–0.87]
DLP	LRP (R ⁺)	0.134 [0.12–0.15]	0.287 [0.27–0.30]	0.155 [0.14–0.18]	0.649 [0.61–0.69]	0.800 [0.71–0.87]	0.609 [0.56–0.65]
DRM	GradCAM	0.513 [0.48–0.54]	0.389 [0.36–0.42]	0.660 [0.62–0.69]	0.668 [0.60–0.73]	0.867 [0.79–0.92]	0.592 [0.53–0.66]
DRM	LRP (R ⁺)	0.548 [0.52–0.58]	0.346 [0.32–0.37]	0.664 [0.63–0.70]	0.640 [0.59–0.69]	0.806 [0.72–0.87]	0.546 [0.49–0.60]
Esophagitis	GradCAM	0.212 [0.20–0.23]	0.215 [0.20–0.23]	0.115 [0.11–0.13]	0.247 [0.21–0.29]	0.300 [0.22–0.40]	0.242 [0.20–0.28]
Esophagitis	LRP (R ⁺)	0.046 [0.04–0.05]	0.206 [0.19–0.22]	0.122 [0.11–0.13]	0.203 [0.17–0.24]	0.140 [0.09–0.22]	0.227 [0.19–0.26]
Cecum	GradCAM	0.128 [0.12–0.13]	0.291 [0.28–0.30]	0.122 [0.12–0.13]	–	–	–
Cecum	LRP (R ⁺)	0.189 [0.18–0.20]	0.333 [0.32–0.35]	0.121 [0.11–0.13]	–	–	–
Pylorus	GradCAM	0.326 [0.31–0.34]	0.403 [0.39–0.41]	0.278 [0.26–0.29]	–	–	–
Pylorus	LRP (R ⁺)	0.105 [0.10–0.11]	0.429 [0.41–0.44]	0.290 [0.27–0.31]	–	–	–
z-line	GradCAM	0.439 [0.42–0.46]	0.498 [0.47–0.52]	0.386 [0.37–0.41]	–	–	–
z-line	LRP (R ⁺)	0.092 [0.09–0.10]	0.436 [0.42–0.45]	0.278 [0.26–0.30]	–	–	–
UC	GradCAM	0.459 [0.44–0.48]	0.466 [0.45–0.48]	0.326 [0.31–0.35]	–	–	–
UC	LRP (R ⁺)	0.529 [0.51–0.55]	0.460 [0.44–0.47]	0.330 [0.31–0.35]	–	–	–

A Tabela 15 reporta comparações pareadas entre GradCAM e LRP (R⁺), estratificadas por modelo (regularizado versus não regularizado), classe e métrica. Para cada cenário, Δ denota a diferença média pareada (GradCAM menos LRP (R⁺)), p_{Holm} é o p-valor corrigido por Holm dentro da família de comparações correspondente, d_z é o tamanho de efeito pareado e n é o número de amostras pareadas. Esta tabela identifica onde as diferenças entre métodos têm suporte estatístico e quantifica sua magnitude, separando evidências baseadas em deleção de fidelidade causal de evidências baseadas em localização de alinhamento com a ROI.

Tabela 15: Comparação Pareada GradCAM vs LRP (R⁺): Δ , p_{Holm} , d_z , n .

MODELO	CLASSE	METRICA	Δ	p_{Holm}	d_z	n
reg	TODAS	PF	0.089	$< 10^{-3}$	0.55	800
reg	TODAS	RP	0.003	0.876	0.05	800
reg	TODAS	IROF	0.004	0.861	0.06	800
reg	TODAS	ALIR	0.079	$< 10^{-3}$	0.60	397
reg	TODAS	PG	0.101	$< 10^{-3}$	0.28	397
reg	TODAS	TKI	0.134	$< 10^{-3}$	0.86	397
reg	Polyps	PF	0.106	$< 10^{-3}$	2.30	100
reg	Polyps	RP	−0.026	$< 10^{-3}$	−0.61	100
reg	Polyps	IROF	−0.024	$< 10^{-3}$	−0.43	100

Continua na Próxima Página

MODELO	CLASSE	METRICA	Δ	p_{Holm}	d_z	n
reg	Polyps	ALIR	0.085	$< 10^{-3}$	0.66	99
reg	Polyps	PG	0.030	0.375	0.14	99
reg	Polyps	TKI	0.263	$< 10^{-3}$	2.18	99
reg	DLP	PF	0.036	$< 10^{-3}$	0.73	100
reg	DLP	RP	-0.001	0.726	-0.03	100
reg	DLP	IROF	-0.025	$< 10^{-3}$	-0.53	100
reg	DLP	ALIR	0.160	$< 10^{-3}$	1.21	100
reg	DLP	PG	0.150	$< 10^{-3}$	0.39	100
reg	DLP	TKI	0.212	$< 10^{-3}$	1.56	100
reg	DRM	PF	-0.035	0.005	-0.28	100
reg	DRM	RP	0.042	$< 10^{-3}$	0.50	100
reg	DRM	IROF	-0.004	0.566	-0.06	100
reg	DRM	ALIR	0.028	0.006	0.22	98
reg	DRM	PG	0.061	0.180	0.16	98
reg	DRM	TKI	0.046	$< 10^{-3}$	0.48	98
reg	Esophagitis	PF	0.167	$< 10^{-3}$	2.42	100
reg	Esophagitis	RP	0.009	$< 10^{-3}$	0.31	100
reg	Esophagitis	IROF	-0.006	0.057	-0.22	100
reg	Esophagitis	ALIR	0.044	$< 10^{-3}$	0.45	100
reg	Esophagitis	PG	0.160	$< 10^{-3}$	0.38	100
reg	Esophagitis	TKI	0.015	0.421	0.15	100
reg	Cecum	PF	-0.061	$< 10^{-3}$	-1.16	100
reg	Cecum	RP	-0.043	$< 10^{-3}$	-1.10	100
reg	Cecum	IROF	0.002	0.267	0.05	100
reg	Pylorus	PF	0.220	$< 10^{-3}$	3.97	100
reg	Pylorus	RP	-0.026	$< 10^{-3}$	-0.50	100
reg	Pylorus	IROF	-0.012	0.024	-0.22	100
reg	z-line	PF	0.346	$< 10^{-3}$	3.80	100
reg	z-line	RP	0.061	$< 10^{-3}$	0.88	100
reg	z-line	IROF	0.109	$< 10^{-3}$	1.41	100
reg	UC	PF	-0.070	$< 10^{-3}$	-0.55	100
reg	UC	RP	0.006	0.140	0.15	100
reg	UC	IROF	-0.003	0.802	-0.05	100

Continua na Próxima Página

MODELO	CLASSE	METRICA	Δ	p_{Holm}	d_z	n
noreg	TODAS	PF	0.268	$< 10^{-3}$	0.98	800
noreg	TODAS	RP	0.024	$< 10^{-3}$	0.21	800
noreg	TODAS	IROF	0.020	$< 10^{-3}$	0.21	800
noreg	TODAS	ALIR	0.128	$< 10^{-3}$	0.78	398
noreg	TODAS	PG	0.254	$< 10^{-3}$	0.50	398
noreg	TODAS	TKI	0.193	$< 10^{-3}$	1.06	398
noreg	Polyps	PF	0.429	$< 10^{-3}$	3.01	100
noreg	Polyps	RP	-0.071	$< 10^{-3}$	-0.95	100
noreg	Polyps	IROF	-0.040	$< 10^{-3}$	-0.58	100
noreg	Polyps	ALIR	0.148	$< 10^{-3}$	0.97	100
noreg	Polyps	PG	0.320	$< 10^{-3}$	0.63	100
noreg	Polyps	TKI	0.336	$< 10^{-3}$	2.19	100
noreg	DLP	PF	0.314	$< 10^{-3}$	1.54	100
noreg	DLP	RP	0.015	0.122	0.19	100
noreg	DLP	IROF	-0.001	0.770	-0.02	100
noreg	DLP	ALIR	0.192	$< 10^{-3}$	1.25	100
noreg	DLP	PG	0.250	$< 10^{-3}$	0.46	100
noreg	DLP	TKI	0.256	$< 10^{-3}$	1.46	100
noreg	DRM	PF	-0.152	$< 10^{-3}$	-0.66	100
noreg	DRM	RP	0.058	$< 10^{-3}$	0.67	100
noreg	DRM	IROF	0.032	0.001	0.39	100
noreg	DRM	ALIR	0.029	0.007	0.28	98
noreg	DRM	PG	0.092	0.078	0.20	98
noreg	DRM	TKI	0.108	$< 10^{-3}$	1.07	98
noreg	Esophagitis	PF	0.304	$< 10^{-3}$	1.92	100
noreg	Esophagitis	RP	0.163	$< 10^{-3}$	1.18	100
noreg	Esophagitis	IROF	0.141	$< 10^{-3}$	1.15	100
noreg	Esophagitis	ALIR	0.142	$< 10^{-3}$	0.74	100
noreg	Esophagitis	PG	0.350	$< 10^{-3}$	0.73	100
noreg	Esophagitis	TKI	0.070	$< 10^{-3}$	0.48	100
noreg	Cecum	PF	-0.019	0.088	-0.20	100
noreg	Cecum	RP	0.032	$< 10^{-3}$	0.46	100
noreg	Cecum	IROF	0.036	$< 10^{-3}$	0.53	100

Continua na Próxima Página

MODELO	CLASSE	METRICA	Δ	p_{Holm}	d_z	n
noreg	Pylorus	PF	0.318	$< 10^{-3}$	3.98	100
noreg	Pylorus	RP	0.023	0.016	0.33	100
noreg	Pylorus	IROF	-0.018	0.001	-0.31	100
noreg	z-line	PF	0.571	$< 10^{-3}$	3.88	100
noreg	z-line	RP	-0.063	$< 10^{-3}$	-0.84	100
noreg	z-line	IROF	-0.024	$< 10^{-3}$	-0.38	100
noreg	UC	PF	0.379	$< 10^{-3}$	2.54	100
noreg	UC	RP	0.032	0.012	0.30	100
noreg	UC	IROF	0.037	$< 10^{-3}$	0.43	100