UNIVERSIDADE FEDERAL DO CARIRI
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PEDRO DA SILVA VIANA

DIAGNÓSTICOS DE ANOMALIAS GASTROINTESTINAIS EM IMAGENS ENDOSCÓPICAS USANDO *ENSEMBLE STACKING* DE CNNS E *VISION TRANSFORMERS*

JUAZEIRO DO NORTE - CE
2026

PEDRO DA SILVA VIANA

DIAGNÓSTICOS DE ANOMALIAS GASTROINTESTINAIS EM IMAGENS ENDOSCÓPICAS USANDO *ENSEMBLE STACKING* DE CNNS E *VISION TRANSFORMERS*

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências e Tecnologia da Universidade Federal do Cariri, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientadora: Prof. Dra. Luana Batista da Cruz

JUAZEIRO DO NORTE - CE

2026

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Cariri
Sistema de Bibliotecas

V614d Viana, Pedro da Silva.

Diagnósticos de anomalias gastrointestinais em imagens endoscópicas usando *ensemble stacking* de *CNNs* e *vision transformers* / Pedro da Silva Viana. – 2026.
59 f. : il. color. – Inclui bibliografia: f. 53-59.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Centro de Ciências e Tecnologia, Departamento de Ciência da Computação, Universidade Federal do Cariri, Juazeiro do Norte, CE, 2026.

Orientadora: Prof. Dra. Luana Batista da Cruz.

1. Imagem Endoscópica. 2. Aprendizagem Profunda. 3. Redes Neurais Convolucionais. 4. Vision Transformers. 5. Ensemble. I. Cruz, Luana Batista da (Orient.). II. Universidade Federal do Cariri. III. Título.

CDD 006.3

Bibliotecário: João Bosco Dumont do Nascimento – CRB 3/1355

PEDRO DA SILVA VIANA

DIAGNÓSTICOS DE ANOMALIAS GASTROINTESTINAIS EM IMAGENS ENDOSCÓPICAS USANDO *ENSEMBLE STACKING* DE CNNS E *VISION TRANSFORMERS*

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências e Tecnologia da Universidade Federal do Cariri, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: 27/03/2026.

BANCA EXAMINADORA

Prof. Dra. Luana Batista da Cruz (Orientadora)
Universidade Federal do Cariri (UFCA)

Prof. Dr. Nelson Carvalho Sandes
Universidade Federal do Cariri (UFCA)

Prof. Dr. João Otávio Bandeira Diniz
Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)

AGRADECIMENTOS

Primeiramente, agradeço a Deus por sempre me conceder forças para seguir em frente, independentemente de quão desafiador e turbulento tenha sido o percurso.

À minha mãe, Nivia Maria da Silva Viana, por ter sido a primeira a acreditar no meu potencial, sempre me encorajando e me sustentando com seu amor e carinho incondicionais. Ao meu pai, Edison Vicente Viana, pelo apoio, pelo amor constante e pelos ensinamentos transmitidos, tanto em seus acertos quanto em seus erros, que serviram de guia para minha formação pessoal. Ao meu irmão, Kauan da Silva Viana, pelas brincadeiras, pelo companheirismo e até mesmo pelos momentos em que conseguiu me tirar do sério, mas que, no fim, sempre contribuíram para o meu crescimento pessoal. A todos os familiares que me apoiaram nesta jornada, expresso minha gratidão pelo incentivo e carinho.

Minha sincera gratidão à minha orientadora, professora Luana Batista da Cruz, pela dedicação e paciência em esclarecer dúvidas, bem como pelo conhecimento compartilhado, que foi fundamental para a realização deste trabalho. Agradeço também a todos os professores da UFCA pelos ensinamentos, conselhos e contribuições ao longo da minha formação. Aos colegas e amigos que fiz na universidade, registro minha gratidão pela amizade, pelo apoio e pelos momentos de descontração, que tornaram essa trajetória muito mais leve e enriquecedora. Em especial, agradeço aos amigos do Go Platypus, que estiveram ao meu lado até mesmo nas ideias mais inusitadas, sempre me apoiando e acreditando junto comigo.

À UFCA, agradeço pela oportunidade de aprendizado e pela estrutura oferecida, essenciais para o desenvolvimento deste trabalho. Por fim, deixo meus agradecimentos a todos aqueles que, de alguma forma, contribuíram e me apoiaram ao longo desse percurso acadêmico.

RESUMO

A detecção automatizada em imagens médicas oferece benefícios significativos, especialmente na identificação precoce de casos clínicos. Este estudo foca em imagens obtidas por exames endoscópicos digestivos, classificando-as em categorias normais (*normal-cecum*, *normal-pylorus* e *normal-z-line*) e anormais, que correspondem a alterações patológicas gastrointestinais (*dyed-lifted-polyps*, *dyed-resection-margins*, *esophagitis*, *polyps* e *ulcerative colitis*). O método proposto engloba extração de Região de Interesse, *Specular Highlights*, *Data Augmentation* e *Ensemble Stacking*, integrando Redes Neurais Convolucionais e *Vision Transformers*. O modelo final alcançou 98,12% de acurácia, 98,15% de precisão, 98,12% de sensibilidade, 98,23% de especificidade e F1-score de 98,13%. Esses resultados indicam que a abordagem proposta tem grande potencial para se tornar uma ferramenta eficaz na identificação de anomalias em exames endoscópicos, contribuindo significativamente para diagnósticos médicos assistidos por inteligência artificial. Além de aumentar a confiabilidade dos diagnósticos, a adoção desse sistema pode otimizar o tempo dos profissionais de saúde, permitindo que se concentrem em casos mais complexos e críticos.

Palavras-chave: Imagem Endoscópica; Aprendizagem Profunda; Redes Neurais Convolucionais; *Vision Transformers*; *Ensemble*.

ABSTRACT

Automated detection in medical images offers significant benefits, especially in the early identification of clinical cases. This study focuses on images obtained from digestive endoscopes examinations, classifying them into normal categories (normal-cecum, normal-pylorus and normal-z-line) and abnormal categories, which correspond to pathological alterations of the gastrointestinal tract (dyed-lifted-polyps, dyed-resection-margins, esophagitis, polyps and ulcerative colitis). The proposed method encompasses Region of Interest extraction, Specular Highlight, Data Augmentation, and Ensemble Stacking, integrating Convolutional Neural Networks and Vision Transformers. The final model achieved 98.12% accuracy, 98.15% precision, 98.12% sensitivity, 98.23% specificity, and a 98.13% F1-score. These results indicate that the proposed approach has great potential to become an effective tool for identifying anomalies in endoscopic examinations, significantly contributing to AI-assisted medical diagnoses. In addition to increasing the reliability of diagnoses, adopting this system can optimize healthcare professionals' time, allowing them to focus on more complex and critical cases.

Keywords: Endoscopy Imaging; Deep Learning; Convolutional Neural Networks; Vision Transformers; Ensemble.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	Acurácia
CAD	*Computer Aided Detection*
CADx	*Computer Aided Diagnosis*
CNNs	*Convolutional Neural Networks*
DA	*Data Augmentation*
DLP	*Dyed Lifted Polyps*
DRM	*Dyed Resection Margins*
ESO	*Esophagitis*
ESP	Especificidade
FN	Falso Negativo
FP	Falso Positivo
IA	Inteligência Artificial
NC	*Normal Cecum*
NP	*Normal Pylorus*
NZL	*Normal Z Line*
P	*Polyps*
PRE	Precisão
RL	Regressão Logística
ROI	Região de Interesse
SEN	Sensibilidade
SH	*Specular Highlights*
UCE	*Ulcerative Colitis*
ViT	*Vision Transformers*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O sistema digestivo é composto por um conjunto de órgãos cuja função primordial é a absorção de nutrientes. A digestão envolve processos físicos e químicos desde a mastigação até a excreção dos resíduos. No entanto, como em qualquer sistema biológico, podem ocorrer anormalidades que afetam seu funcionamento. Estas anormalidades podem variar desde condições relativamente benignas até doenças graves que requerem intervenção médica. Dois exemplos notáveis são os pólipos e a colite ulcerativa (ROGLER, 2014; CHIRAS, 2013).

Os pólipos são pequenos crescimentos anormais que podem se assemelhar a cogumelos e são encontrados em diferentes regiões do corpo, como o estômago e o intestino. Embora sejam considerados tumores benignos, sua identificação é crucial, pois seu surgimento pode indicar a presença de câncer na área em que se encontram (NOFFSINGER, 2008). A colite ulcerativa é uma inflamação intestinal presente no cólon, que pode aumentar as chances de desenvolvimento de câncer. Identificar essas anormalidades o mais cedo possível é extremamente benéfico, pois pode ser essencial para o diagnóstico precoce de câncer gastrointestinal, como o câncer de estômago e o câncer colorretal (ROGLER, 2014).

O câncer de estômago é uma das neoplasias malignas mais incidentes no Brasil, ocupando a quinta posição em termos de ocorrência, segundo (INCA, 2022). Trata-se de uma enfermidade agressiva, que acomete principalmente indivíduos acima de 50 anos e pessoas com obesidade. De forma semelhante, o câncer colorretal também apresenta alta relevância epidemiológica. De acordo com (INCA, 2022), é a terceira neoplasia mais comum no país, com maior prevalência em indivíduos com predisposição genética, obesidade e consumo excessivo de álcool.

A detecção precoce é crucial para o tratamento eficaz e a melhora do prognóstico dos pacientes. A identificação de anormalidades no trato digestivo é realizada por meio de procedimentos principais, considerados padrão-ouro: a endoscopia digestiva alta e a colonoscopia. A endoscopia digestiva alta consiste na introdução de um endoscópio pela boca do paciente, permitindo ao médico visualizar diretamente a mucosa do esôfago, estômago e duodeno por meio de imagens (ILIC; ILIC, 2022). Já a colonoscopia segue o mesmo princípio, mas o endoscópio é inserido pelo ânus do paciente para visualizar a porção inferior do intestino, como o cólon (KOH *et al.*, 2021).

Um dos desafios significativos enfrentados pelos endoscopistas é a quantidade massiva de dados gerados durante a endoscopia digestiva alta. Essa sobrecarga de informações

pode ser otimizada com o auxílio de sistemas computacionais como Detecção Assistida por Computador (*Computer Aided Detection* - CAD) e Diagnóstico Assistido por Computador (*Computer Aided Diagnosis* - CADx). Esses sistemas dão suporte na análise de doenças, acelerando e aprimorando o processo de análise e diagnóstico (ALAGAPPAN *et al.*, 2018). É possível observar na literatura uma variedade de métodos computacionais desenvolvidos para detecção e diagnóstico de doenças por meio de imagens médicas (CRUZ *et al.*, 2020; JÚNIOR *et al.*, 2021; CRUZ *et al.*, 2022; DINIZ *et al.*, 2023). Nesse contexto, métodos computacionais tornam-se ferramentas promissoras para apoiar especialistas na análise de exames endoscópicos.

1.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um método computacional capaz de classificar automaticamente os exames de colonoscopia e endoscopia digestiva alta em duas categorias: normais e anormais. Para isso, emprega-se uma estratégia de *Ensemble Stacking* composta por modelos baseados em *Convolutional Neural Networks* (CNNs) e *Vision Transformers*, especificamente EfficientNet-B4, ResNet-50 e PVTv2-B2, em conjunto com técnicas de pré-processamento, como extração da Região de Interesse (ROI), redução de *Specular Highlights* (SH) e *Data Augmentation* (DA).

1.2 Objetivos Específicos

A fim de atingir o objetivo geral delineado neste trabalho, foram estabelecidos os seguintes objetivos específicos:

- Realizar o levantamento da literatura para compreender o problema e identificar lacunas de pesquisa;
- Realizar a extração da ROI das imagens, de modo a reduzir informações irrelevantes e otimizar o processo de treinamento;
- Implementar técnicas de DA a fim de aumentar a diversidade do conjunto de treinamento;
- Investigar e aplicar CNNs e *Vision Transformers* visando à realização da tarefa de classificação anomalias gastrointestinais;
- Investigar e aplicar métodos *Ensemble* para identificar a melhor estratégia na tarefa de classificação de anomalias gastrointestinais;
- Validar o método proposto utilizando métricas de desempenho amplamente empregadas

em trabalhos de análise de imagens médicas;

- Investigar a contribuição de cada etapa do método proposto por meio da realização de estudos de ablação e testes de hipótese;
- Comparar o desempenho do método proposto com abordagens já consolidadas na literatura, de forma a contextualizar os resultados e demonstrar sua contribuição.

1.3 Organização do Trabalho

Os capítulos restantes que compõem este trabalho estão organizados da seguinte forma:

- Capítulo 2 apresenta uma revisão da literatura, reunindo os principais trabalhos relacionados à tarefa de classificação de anormalidades gastrointestinais em imagens endoscópicas.
- Capítulo 3 apresenta a fundamentação teórica que embasa a pesquisa, detalhando os principais conceitos utilizados, como as arquiteturas de *Deep Learning* empregadas, os testes de hipótese e as estratégias de *Ensemble Learning* utilizadas na classificação.
- Capítulo 4 descreve as etapas do método proposto para a classificação das classes normais e anormais, contemplando o base de imagens utilizado, a extração da ROI, o pré-processamento das imagens, o *Ensemble Stacking*, métricas de validação e o teste de hipótese.
- Capítulo 5 expõe e discute os resultados obtidos a partir dos experimentos realizados, analisando o desempenho do método proposto com base em métricas de avaliação. Além de apresentar estudos de caso e uma análise comparativa com os trabalhos relacionados, destacando as vantagens da abordagem desenvolvida.
- Capítulo 6 traz as considerações finais, sintetizando as principais contribuições deste trabalho e apresentando direções para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

As seções seguintes apresentam os conceitos fundamentais relacionados aos exames de colonoscopia e endoscopia digestiva alta, à anatomia do esôfago, estômago e do intestino grosso, bem como às principais anormalidades gastrointestinais. Também são discutidos os fundamentos de Inteligência Artificial (IA), *Machine Learning* e *Deep Learning*, com ênfase em Rede Neural Convolucional (*Convolutional Neural Networks* - CNNs) e *Vision Transformers* (ViT), nas arquiteturas empregadas e nas técnicas de Aumento de Dados (*Data Augmentation* - DA). Por fim, são descritas as métricas de desempenho utilizadas na validação dos resultados experimentais.

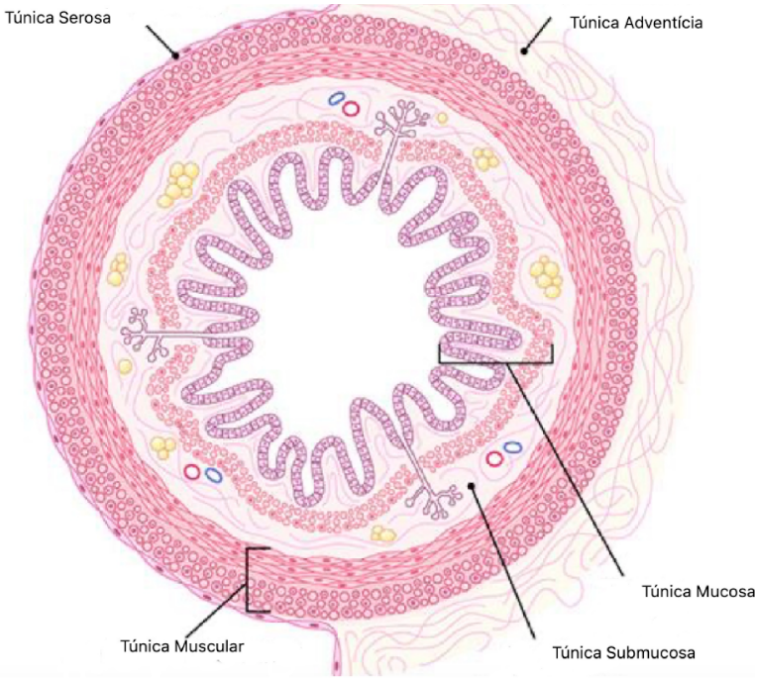
2.1 Esôfago e Esofagite

O esôfago é um tubo muscular alongado que conecta a faringe ao estômago. Apresenta coloração pálida e está localizado na região cervical e torácica do mediastino (região central do tórax), atravessando o diafragma até a porção abdominal superior. Em adultos, o comprimento médio é de aproximadamente 25 cm, variando conforme a estatura e a idade. Na porção proximal (parte de um membro ou órgão que fica mais próxima do centro do corpo ou do seu ponto de origem), encontra-se o esfíncter esofágico superior (válvula muscular no início do esôfago cuja função é liberar a passagem do alimento e impedir que ele vá para os pulmões), enquanto a porção distal termina no esfíncter esofágico inferior (válvula muscular entre o esôfago e o estômago que impede o retorno do alimento e do ácido gástrico), cuja posição em relação ao diafragma é importante para a proteção contra o refluxo gastroesofágico (CHAUDHRY; BORDONI, 2023).

Macroscopicamente, o esôfago segue a organização geral do tubo digestório, constituído por camadas concêntricas, mucosa, submucosa, muscular própria e uma camada externa. Essa camada externa varia de acordo com a localização do órgão, sendo adventícia nos segmentos situados no tórax e pescoço, onde o esôfago se encontra fixado a estruturas adjacentes, e serosa apenas em sua porção abdominal, que é revestida pelo peritônio, como ocorre nos demais órgãos do trato gastrointestinal localizados na cavidade abdominal (estômago, intestino delgado e intestino grosso) (CHAUDHRY; BORDONI, 2023), como ilustrado na Figura 1. A mucosa é revestida por epitélio estratificado, adaptado ao transporte de alimentos, enquanto a camada muscular garante o mecanismo peristáltico, responsável pelo deslocamento do bolo

alimentar. Sua principal função é conduzir e coordenar o transporte de alimentos e líquidos da faringe ao estômago, preservando a integridade da mucosa frente a agentes físicos e químicos (CHAUDHRY; BORDONI, 2023).

Figura 1 – Organização geral do tubo digestório.



Fonte: Adaptado de (UNIFAL-MG – Histologia Interativa, n.d.).

Entre as alterações que acometem o esôfago, destaca-se a Esofagite (*Esophagitis* – ESO), definida como inflamação da mucosa esofágica. Suas principais causas incluem refluxo gastroesofágico, infecções, reações alérgicas e lesões mecânicas ou químicas. Essa condição pode gerar sintomas como azia, dor retroesternal e disfagia, além de, em casos mais graves, sangramento ou perfuração. Quando persistente, a inflamação aumenta o risco de complicações, incluindo estenose (estreitamento anormal de um canal, vaso sanguíneo ou órgão tubular do corpo), ulceração (formação de uma ferida aberta na pele ou na membrana mucosa de um órgão) e metaplasia (substituição de um tipo de célula por outro diferente) (DELLON *et al.*, 2025).

O diagnóstico da ESO é realizado, principalmente, por meio da endoscopia digestiva alta, que permite a inspeção direta, confirmação da ESO, determinação da causa e avaliação de displasia ou neoplasia (SIMADIBRATA *et al.*, 2023). O diagnóstico precoce dessas alterações, incluindo lesões precursoras de câncer, melhora significativamente o prognóstico.

2.2 Estômago e Pólipos

O estômago é um órgão sacular (órgão em formato de saco, que tem a função de armazenar substâncias temporariamente) e muscular localizado no hipocôndrio esquerdo e na região epigástrica, desempenhando a função de armazenar, misturar e fragmentar os alimentos para que o intestino delgado possa absorver os nutrientes. Nesse processo, forma-se o quimo, uma pasta semi-líquida resultante da mistura de alimentos e sucos digestivos (FEUERSTEIN *et al.*, 2019). A parede gástrica mantém a mesma organização estrutural geral do tubo digestório, descrita na Seção 2.1 e ilustrada na Figura 1.

Entre as alterações que podem acometer a mucosa gástrica, destacam-se os Pólipos (*Polyps* – P), definidos como lesões elevadas decorrentes do crescimento anormal do tecido epitelial. Sua formação pode estar associada a fatores como gastrite crônica, infecções e predisposição genética (ISLAM NEAL C. PATEL; NGUYEN, 2014). Além do estômago, pólipos também podem ocorrer no intestino grosso, relacionados a fatores semelhantes, com exceção da gastrite, que não influencia sua formação nessa região (NOFFSINGER, 2008). Embora a maioria seja considerada inofensiva, alguns pólipos apresentam potencial de transformação maligna, podendo evoluir para neoplasias, como câncer gástrico ou colorretal, a depender de sua localização. Quando presentes, os sintomas mais comuns incluem dor abdominal, náuseas, vômitos, saciedade precoce e episódios de sangramento gastrointestinal (ISLAM NEAL C. PATEL; NGUYEN, 2014; NOFFSINGER, 2008).

A detecção precoce dos P por meio da endoscopia é fundamental para prevenir a progressão da doença. O uso de corantes endoscópicos pode facilitar a visualização dessas lesões, gerando imagens de Pólipos Levantados com Corante (*Dyed Lifted Polyps* – DLP), e Margens de Ressecção Tingidas (*Dyed Resection Margins* – DRM), tornando o diagnóstico mais ágil e preciso, além de facilitar a remoção das do P. Essa estratégia tem grande relevância clínica, pois a rápida classificação dos P contribui significativamente para a prevenção e o tratamento precoce do câncer gástrico e do câncer colorretal (FEUERSTEIN *et al.*, 2019).

2.3 Intestino Grosso e Colite Ulcerativa

O intestino grosso é um órgão tubular que conecta o íleo ao canal anal e desempenha funções essenciais, como a absorção de água e eletrólitos, a formação das fezes e a propulsão do conteúdo fecal em direção ao reto (AZZOUZ; SHARMA, 2023; RAO; WANG, 2010). A parede

do intestino grosso apresenta a mesma organização estrutural geral do tubo digestório, conforme descrito na Seção 2.1 e ilustrado na Figura 1.

A Colite Ulcerativa (*Ulcerative Colitis* – UCE). é uma doença inflamatória intestinal crônica caracterizada por inflamação contínua e superficial da mucosa do intestino grosso, geralmente iniciada no reto e com possível extensão para outras regiões. Embora sua etiologia não esteja totalmente esclarecida, acredita-se que resulte da interação entre predisposição genética, alterações na resposta imunológica, fatores ambientais e desequilíbrios da microbiota intestinal. Suas manifestações clínicas mais comuns incluem diarreia sanguinolenta, tenesmo, urgência evacuatória e dor abdominal. A persistência da atividade inflamatória aumenta o risco de complicações agudas e de consequências crônicas, como displasia e câncer colorretal em casos de longa duração da doença (UNGARO *et al.*, 2017).

O diagnóstico é feito pela correlação de achados clínicos, endoscópicos e histopatológicos, sendo a colonoscopia com biópsia o método mais indicado. Esse exame possibilita a inspeção direta da mucosa, a identificação do padrão de distribuição da inflamação e a coleta de fragmentos para análise histológica, essenciais para confirmar a UCE e diferenciá-la de outras causas de diarreia inflamatória (RUBIN *et al.*, 2019). Em aplicações de processamento de imagem médica, a análise automatizada de exames endoscópicos ou tomográficos pode classificar a gravidade da inflamação, diferenciar lesões benignas de suspeitas de malignidade e indicar áreas prioritárias para biópsia. Esses recursos apoiam o diagnóstico, o monitoramento da resposta terapêutica e o desenvolvimento de ferramentas de triagem e suporte à decisão baseadas em IA (GE *et al.*, 2023). Nesse contexto, no Capítulo 4 demonstra o método utilizado para classificar tais anormalidades em exames endoscópicos, contribuindo para a detecção e análise automatizada dessas condições.

2.4 Endoscopia

A endoscopia fornece imagens em tempo real da superfície mucosa e permite intervenções terapêuticas imediatas, servindo ao endoscopista como base para identificar diversas anormalidades e definir a melhor estratégia terapêutica para cada paciente (AHLAWAT *et al.*, 2023). A Figura 2 é um exemplo da máquina utilizada na realização do exame.

A endoscopia é um exame minimamente invasivo que utiliza um tubo flexível iluminado (endoscópio) com uma câmera na extremidade para visualizar internamente o trato digestivo. O endoscópio integra um sistema de iluminação, câmera e um canal de trabalho

Figura 2 – Exemplo de uma máquina de endoscopia.

Fonte: (Hospicenter, n.d.)

que permite a passagem de instrumentos (pinças de biópsia, laços de polipectomia, sondas hemostáticas), irrigação e aspiração. As imagens captadas são transmitidas em tempo real para um monitor e podem ser gravadas como *frames* ou vídeos digitais para documentação e análise posterior. A concepção e o design dos endoscópios, bem como a ergonomia e os detectores de imagem, explicam a qualidade atual das imagens endoscópicas e as capacidades terapêuticas do equipamento.

Antes da realização do exame, o paciente passa por avaliação e preparo específicos, incluindo jejum, revisão do uso de anticoagulantes, triagem de risco anestésico e planejamento da sedação ou analgesia de acordo com as comorbidades. Essas etapas são obrigatórias tanto para a endoscopia digestiva alta quanto para a colonoscopia. No caso da colonoscopia, acrescenta-se

ainda o preparo intestinal com laxantes. Durante o exame, o paciente é monitorado e sedado conforme protocolos estabelecidos, sendo o tipo e a profundidade da sedação definidos de acordo com a complexidade do procedimento e as diretrizes clínicas de sedação em endoscopia. Tais medidas visam reduzir riscos e garantir maior segurança e conforto ao paciente (AHLAWAT *et al.*, 2023).

A posição do paciente e a imobilização adequada durante a endoscopia impactam a exposição das lesões e a segurança do procedimento. O decúbito lateral esquerdo é usual para colonoscopia, enquanto a posição supina ou com ajuste de inclinação pode ser adotada em procedimentos superiores. Escolhas de posicionamento podem facilitar a visualização e o tratamento de algumas lesões. Além disso, recomenda-se preparação e protocolos de recuperação para monitorar efeitos adversos relacionados à sedação e às intervenções realizadas (KANG; HYUN, 2013).

Em suma, a endoscopia é uma ferramenta diagnóstica e terapêutica central nas doenças do trato digestivo, pois fornece visualização direta e documentada da mucosa, permite coleta de material para histologia e possibilita intervenções imediatas. As melhorias em design de endoscópios, em técnicas de realce de imagem e em protocolos de sedação têm ampliado a eficácia e a segurança das práticas endoscópicas, tornando-as essenciais no planejamento e na execução do tratamento gastroenterológico.

2.5 Pré-processamento

A etapa de pré-processamento desempenha um papel fundamental em sistemas de visão computacional, especialmente na análise de imagens médicas. Segundo (GONZALEZ; WOODS, 2010), o pré-processamento engloba um conjunto de técnicas cuja finalidade principal é aprimorar a qualidade dos dados visuais para torná-los mais adequados ao processamento computacional subsequente. O objetivo central não é realizar inferências sobre a imagem, mas sim prepará-la por meio da redução de ruídos e realce de características estruturais relevantes.

No domínio de exames endoscópicos, a aplicação dessas técnicas pode ser benéfica para garantir que os modelos de aprendizado profundo concentrem-se nos padrões anatômicos e patológicos corretos. Essa fase atua diretamente na redução de artefatos inerentes ao processo de captura, como reflexos intensos gerados pela iluminação do equipamento sobre a mucosa úmida, e na delimitação das áreas úteis da imagem. Dessa forma, o pré-processamento pode fornecer representações mais limpas e consistentes, otimizando a capacidade de generalização e

a precisão das etapas posteriores de classificação.

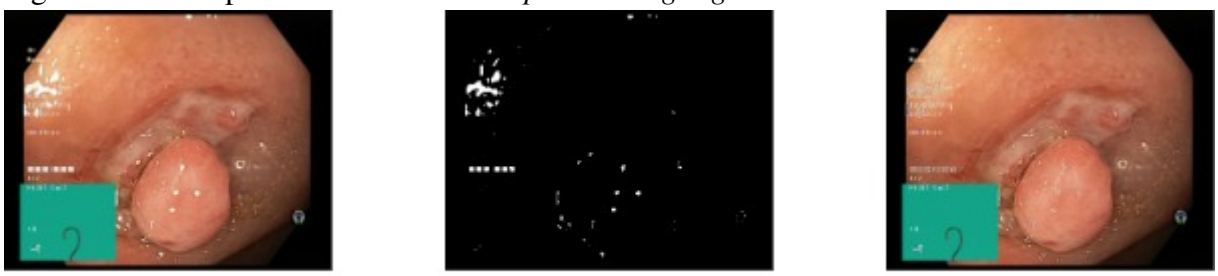
2.5.1 Specular Highlights

Os *Specular Highlights* (reflexos especulares) são focos de brilho intenso causados pela reflexão direta da luz sobre superfícies úmidas, um fenômeno inerente à iluminação de procedimentos endoscópicos (AJLAN *et al.*, 2019). Nas imagens capturadas, esses artefatos saturam os sensores e ocultam as reais texturas e estruturas anatômicas do tecido.

Para os modelos de *Machine Learning*, esses reflexos representam um ruído visual severo. As redes tendem a interpretar o alto brilho e as bordas marcadas desses artefatos como características relevantes. Isso distrai o modelo dos padrões reais da imagem, prejudicando o processo de aprendizado e reduzindo significativamente a capacidade de generalização e a precisão preditiva do classificador (ZHANG *et al.*, 2023).

A atenuação algorítmica desse problema beneficia a modelagem computacional ao restaurar a integridade visual da imagem. A Figura 3 ilustra uma imagem antes e depois do método e a máscara gerada. Essa correção fornece aos modelos dados mais limpos e consistentes, garantindo que o aprendizado ocorra exclusivamente sobre os padrões reais do tecido (ZHANG *et al.*, 2023).

Figura 3 – Exemplo de tratamento dos *Specular Highlights*.



Fonte: Elaborado pelo autor.

2.6 *Data Augmentation*

O DA é uma técnica utilizada para ampliar artificialmente o conjunto de dados por meio da aplicação de transformações nas amostras originais. Seu objetivo é aumentar a variabilidade dos dados disponíveis, contribuindo para o treinamento de modelos mais robustos e com maior capacidade de generalização (SHORTEN; KHOSHGOFTAAR, 2019).

As técnicas de DA podem variar desde operações básicas, como rotações, translações, inversões horizontais e ajustes de brilho e contraste, até métodos mais avançados, como recortes aleatórios e mistura de imagens (SHORTEN; KHOSHGOFTAAR, 2019). Cada técnica é escolhida de acordo com as particularidades da base de dados e do problema em questão, permitindo ao modelo ser exposto a diferentes variações e, consequentemente, tornando-o mais robusto.

Além de aumentar o volume de dados disponíveis, o DA também atua como um processo de regularização, ao introduzir variações artificiais controladas nas amostras, o que reduz a tendência da rede a memorizar padrões específicos e a incentiva a aprender representações mais robustas e generalizáveis. Isso é especialmente relevante em aplicações biomédicas, onde o número de amostras é limitado e a diversidade de casos clínicos é crítica para a eficácia dos modelos de classificação (PEREZ; WANG, 2017).

2.7 Inteligência Artificial

O campo da Inteligência Artificial (IA) começou a ser estruturado a partir da metade do século XX, quando surgiram as primeiras reflexões teóricas e propostas práticas sobre a possibilidade de máquinas apresentarem comportamento inteligente. Em 1950, Alan Turing, no ensaio seminal *Computing Machinery and Intelligence*, propôs um enquadramento conceitual para questionar se máquinas poderiam exibir comportamento inteligente, introduzindo o "*Imitation Game*" como método de avaliação, o que se consolidou como marco teórico inicial para pesquisas subsequentes (TURING, 1950).

A formalização do campo como área de investigação ocorreu em 1956, com o *Dartmouth Summer Research Project on Artificial Intelligence*, organizado por John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon, ocasião em que o termo *artificial intelligence* foi cunhado e um programa interdisciplinar de pesquisa foi delineado, estimulando estudos pioneiros em raciocínio simbólico e resolução de problemas (MCCARTHY *et al.*, 1956). Nesse mesmo período, trabalhos como os de Shannon sobre jogos e representação do conhecimento, bem como o desenvolvimento do *Logic Theorist* por Newell, Shaw e Simon, demonstraram que atividades cognitivas poderiam ser formalizadas e executadas por sistemas computacionais, estabelecendo os primeiros alicerces práticos da IA (SHANNON, 1950; NEWELL *et al.*, 1956). Esses esforços fundadores consolidaram, ao mesmo tempo, os objetivos conceituais de simular percepção e raciocínio humanos e as primeiras estratégias para alcançá-los.

A partir desses marcos fundadores da IA, seus métodos começaram a ser aplicados em diferentes domínios científicos, incluindo a medicina. As primeiras iniciativas voltadas à análise de imagens médicas datam da década de 1960, quando estudos pioneiros demonstraram que computadores poderiam auxiliar no diagnóstico a partir de radiografias. Um exemplo notável é o trabalho de (LODWICK *et al.*, 1963), que descreveu um sistema para apoio no diagnóstico diferencial de tumores ósseos.

Nas décadas seguintes, o campo expandiu-se com aplicações em lesões mamárias e outros esforços de Detecção Assistida por Computador (*Computer Aided Detection* - CAD) durante os anos 1970 e 1980 (ACKERMAN *et al.*, 1972), estabelecendo as bases metodológicas para detecção automática e classificação em imagens médicas. Embora esses sistemas iniciais fossem relativamente simples quando comparados aos modelos atuais, eles foram fundamentais para a consolidação do Diagnóstico Assistido por Computador (*Computer Aided Diagnosis* - CADx) e pavimentaram o caminho para a adoção de técnicas de *Machine Learning* e, mais recentemente, de *Deep Learning* em radiologia e endoscopia (DOI, 2009; AVANZO *et al.*, 2024).

2.7.1 Machine Learning

O *Machine Learning* surgiu como um desdobramento da IA, buscando desenvolver métodos que permitissem aos computadores aprender a partir de dados e melhorar seu desempenho em tarefas específicas sem a necessidade de programação explícita. O termo foi popularizado por Arthur Samuel em 1959, ao descrever programas capazes de aperfeiçoar estratégias em jogos de damas com base na experiência acumulada (SAMUEL, 1959).

Durante as décadas de 1960 e 1970, os primeiros algoritmos de *Machine Learning* foram consolidados, incluindo métodos de regressão, árvores de decisão e perceptrons, que introduziram o paradigma de aprendizado supervisionado (ROSENBLATT, 1958; QUINLAN, 1986). Paralelamente, abordagens não supervisionadas, como o algoritmo *K-Means* (MACQUEEN, 1965), mostraram-se úteis para análise exploratória de dados e agrupamento automático. Esses avanços permitiram que problemas inicialmente restritos ao raciocínio simbólico fossem tratados de forma estatística e adaptativa.

Nos anos 1980 e 1990, a consolidação de técnicas como redes neurais multicamadas treinadas com retropropagação do erro (RUMELHART *et al.*, 1986), métodos baseados em instâncias (*k-nearest neighbors*) (COVER; HART, 1967) e máquinas de vetores de suporte (CORTES; VAPNIK, 1995) expandiu o alcance do campo, especialmente em tarefas de classificação

e reconhecimento de padrões. A crescente disponibilidade de dados digitais e o aumento da capacidade computacional fortaleceram esse movimento, tornando o *Machine Learning* uma disciplina central para aplicações práticas da IA.

2.7.2 Deep Learning

O *Deep Learning* é um ramo do *Machine Learning* que explora arquiteturas baseadas em redes neurais artificiais com múltiplas camadas. Consolidado no século XXI, esse paradigma tem impulsionado avanços significativos em áreas como visão computacional, processamento de linguagem natural e análise de imagens (LECUN *et al.*, 2015). Seu funcionamento busca emular, mecanismos de processamento do cérebro humano, permitindo que modelos extraiam representações hierárquicas e de alta complexidade a partir dos dados (LECUN *et al.*, 2015). Essa perspectiva é inspirada em estudos neurofisiológicos de (HUBEL; WIESEL, 1998), que evidenciaram a organização hierárquica do córtex visual, com diferentes camadas especializadas no processamento de estímulos visuais.

Essa abordagem tem sido aplicada com sucesso em áreas centrais da IA, como processamento de linguagem natural, visão computacional, análise semântica e transferência de aprendizado. Seu crescimento acelerado foi impulsionado por três fatores principais, sendo eles o aumento da capacidade de processamento das unidades gráficas, o aprimoramento dos algoritmos de treinamento e a significativa redução no custo do hardware (GUO *et al.*, 2016).

Diferentemente das redes neurais artificiais tradicionais, as técnicas de *Deep Learning* permitem a extração automática de características diretamente dos dados, reduzindo etapas manuais de pré-processamento e engenharia de atributos. Com múltiplas camadas não lineares, essas arquiteturas são capazes de realizar abstrações sucessivas que integram extração, seleção e classificação de características em um único modelo, reduzindo a necessidade de intervenção humana (LECUN *et al.*, 2015; HUA *et al.*, 2015).

As pesquisas em *Deep Learning* têm produzido um conjunto variado de arquiteturas, cada uma desenvolvida para explorar características específicas dos dados e das tarefas. Entre as mais conhecidas estão as CNNs, que se destacam em aplicações de visão computacional, e as Redes Neurais Recorrentes, frequentemente utilizadas em problemas de natureza sequencial, como séries temporais e processamento de linguagem natural. Modelos mais avançados, como as Redes de Memória de Longo Prazo, surgiram como uma solução às limitações das Redes Neurais Recorrentes tradicionais, enquanto estruturas como *autoencoders* empilhados e Redes de Crença

Profunda possibilitaram novos avanços em aprendizado não supervisionado. O ecossistema reflete a rápida evolução do campo, com inovações que expandem aplicações do *Deep Learning* e o tornam mais próximo do desempenho humano (SHRESTHA; MAHMOOD, 2019).

2.7.2.1 Redes Neurais Convolucionais

As CNNs são arquiteturas, capazes de aprender representações hierárquicas de dados estruturados em matrizes, como imagens, vídeos e sinais temporais (LECUN *et al.*, 2015). Essa capacidade decorre de princípios fundamentais, incluindo conexões locais, compartilhamento de pesos, camadas de *pooling* e empilhamento de múltiplas camadas, que permitem que a rede capture padrões cada vez mais complexos à medida que os dados avançam pela arquitetura.

O funcionamento das CNNs se baseia principalmente em camadas convolucionais e de *pooling*. As camadas convolucionais aplicam filtros locais para detectar características simples, como bordas e texturas, que, em camadas mais profundas, são combinadas em representações hierárquicas mais abstratas. As camadas de *pooling* reduzem a dimensionalidade e conferem invariância a pequenas variações na posição ou na forma dos padrões detectados, tornando a rede mais robusta e eficiente (LECUN *et al.*, 2015). Essa combinação de operações permite que as CNNs aprendam diretamente dos dados, minimizando a necessidade de extração manual de características.

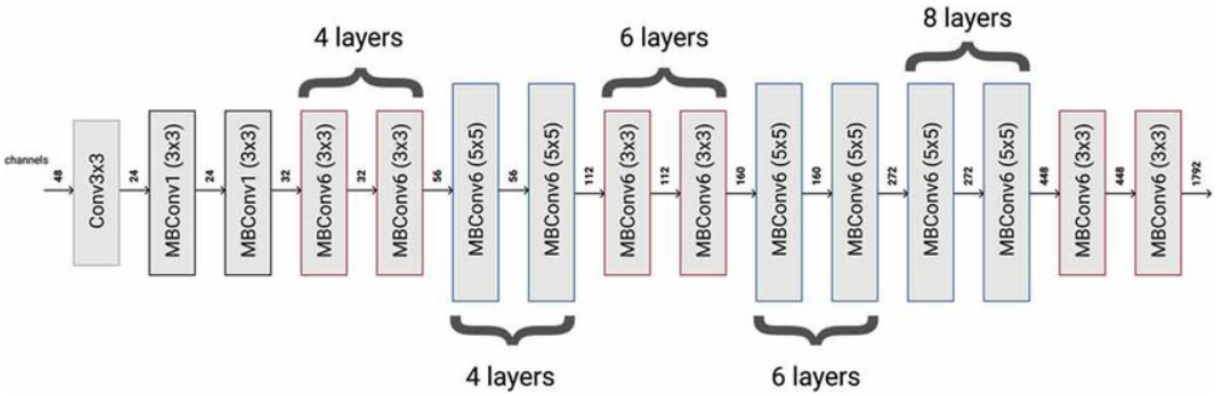
Historicamente, as CNNs tiveram impacto significativo em diversas aplicações práticas, desde reconhecimento de fala e leitura de documentos nos anos 1990 até sistemas modernos de visão computacional, detecção de objetos e reconhecimento facial (LECUN *et al.*, 1998). Inspiradas na neurociência visual, essas redes se consolidaram como um dos pilares do *Deep Learning*, sendo amplamente utilizadas em problemas que envolvem dados com estrutura espacial ou temporal, e continuam a evoluir com novas arquiteturas e técnicas de treinamento (SHRESTHA; MAHMOOD, 2019).

2.7.2.1.1 Arquitetura EfficientNet-B4

As EfficientNets, propostas por (TAN; LE, 2019), formam uma família de arquiteturas convolucionais desenhadas para maximizar a acurácia mantendo eficiência computacional. Elas combinam busca automática por arquitetura (*neural architecture search*) com o princípio de *compound scaling*, que escala de forma balanceada a largura, a profundidade e a resolução da rede em vez de aumentar apenas um desses fatores. A família emprega blocos MBConv

com mecanismos de *squeeze-and-excitation* e utiliza funções de ativação como swish/SiLU, o que resulta em boa reutilização de parâmetros e numa favorável relação entre acurácia e custo computacional. As variantes, por exemplo, B0–B7 permitem escolher diferentes *trade-offs* entre desempenho e requisitos de recurso. A Figura 4 ilustra a arquitetura da EfficientNet-B4.

Figura 4 – Arquitetura EfficientNet-B4.



Fonte: (PAK *et al.*, 2020).

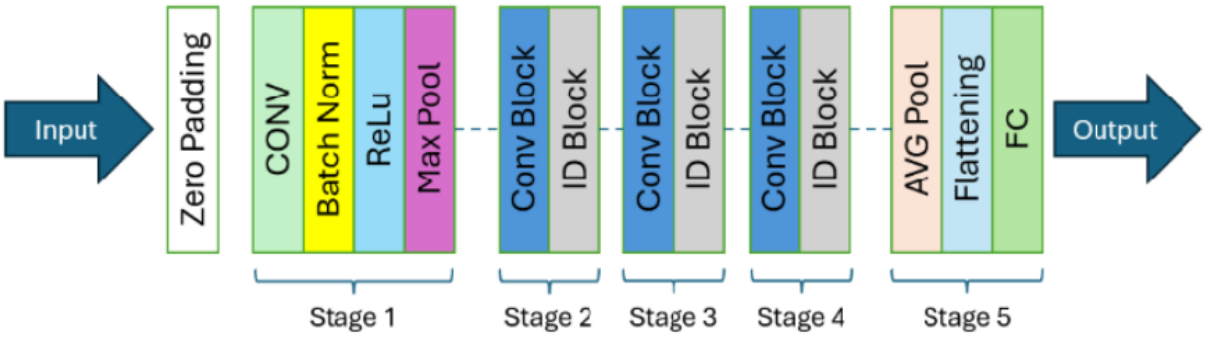
A EfficientNet-B4 é gerada pela aplicação sistemática do esquema de *compound scaling* sobre a arquitetura original. Enquanto o modelo base foi concebido para ser extremamente leve, a versão B4 expande a largura, a profundidade e a resolução de entrada de forma coordenada para capturar padrões espaciais mais complexos. Essa configuração intermediária entrega um excelente equilíbrio entre um alto desempenho preditivo e viabilidade computacional. Devido a essa capacidade ampliada de extração de características sem o custo proibitivo das versões mais profundas, a EfficientNet-B4 atua de forma robusta como *backbone* em tarefas sensíveis e densas de visão computacional, como a classificação de exames endoscópicos, sendo também uma excelente base para *transfer learning* (TAN; LE, 2019).

2.7.2.1.2 Arquitetura ResNet-50

As ResNets, propostas por (HE *et al.*, 2015), constituem uma família de arquiteturas concebidas para reduzir a degradação de desempenho quando as redes se tornam muito profundas. O princípio-chave é a introdução de conexões residuais, atalhos que somam a entrada de um bloco à sua saída, permitindo que as camadas aprendam funções residuais em vez de mapear diretamente transformações complexas. Esse esquema estabiliza o fluxo de gradiente, facilita o treinamento de redes com dezenas ou centenas de camadas e torna possível empregar profundidades maiores sem perda de desempenho. Existem várias variantes da família, por exemplo, ResNet-18, ResNet-34, ResNet-50, que diferem em profundidade e no tipo de blocos usados,

mas preservam o mesmo conceito de conectividade residual A Figura 5 ilustra a arquitetura típica da ResNet-50.

Figura 5 – Arquitetura ResNet-50.



Fonte: (RISKA *et al.*, 2025).

A ResNet-50 é uma instância amplamente utilizada dessa família, composta por 50 camadas organizadas em blocos residuais do tipo *bottleneck*. Cada bloco *bottleneck* tipicamente combina três convoluções sequenciais: 1×1 para reduzir dimensões, 3×3 para processamento espacial e 1×1 para restaurar dimensões, e utiliza atalhos de identidade ou atalhos com projeção (1×1) quando é necessária alteração de dimensão. Essa arquitetura mantém estabilidade no treinamento, melhora a capacidade de generalização e oferece um bom compromisso entre profundidade e custo computacional, por isso é frequentemente empregada como *backbone* em tarefas de visão computacional (HE *et al.*, 2015).

2.7.2.2 Transformers e Vision Transformers

A arquitetura *Transformer*, introduzida inicialmente para tarefas de processamento de linguagem natural por (VASWANI *et al.*, 2017), representa um marco no *Deep Learning* ao substituir a recorrência e as convoluções tradicionais pelo mecanismo de autoatenção (*self-attention*). Esse mecanismo permite que o modelo pondere dinamicamente a importância de diferentes partes dos dados de entrada, independentemente da distância espacial ou temporal entre elas. Essa característica possibilita o processamento paralelo das informações e uma captura extremamente eficiente de dependências de longo alcance, servindo de base para o desenvolvimento de novos modelos em diversas áreas da IA.

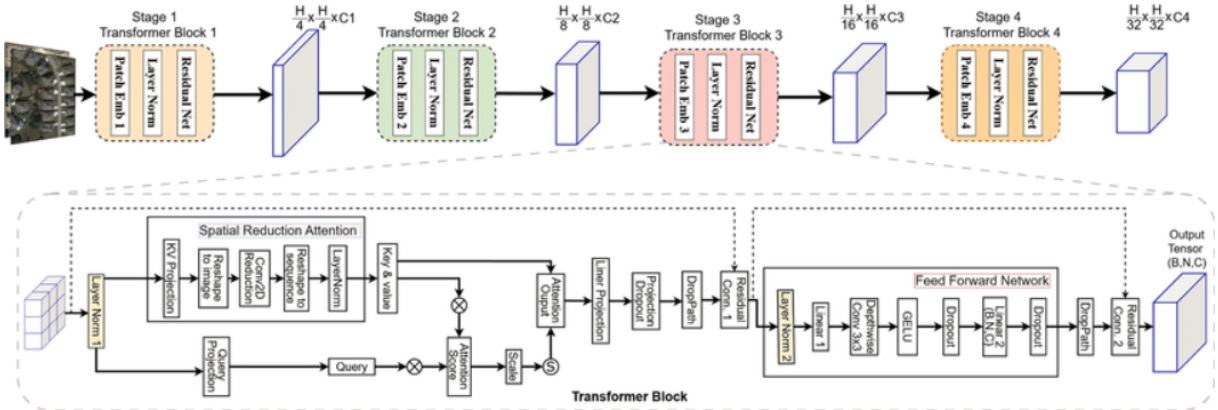
Os *Vision Transformers* (ViT), propostos por (DOSOVITSKIY *et al.*, 2021), adaptaram a arquitetura *Transformer* para o domínio da visão computacional. Em vez de utilizar filtros convolucionais locais, o ViT divide a imagem original em uma grade de pequenos blocos de

tamanho fixo (*patches*), tratando cada bloco de forma análoga a uma palavra em um texto. Esses *patches* são achatados em vetores, recebem informações sobre sua localização original e são processados pelas camadas de atenção. Ao contrário das CNNs, que constroem o entendimento global da imagem de forma gradual e hierárquica, os ViTs são capazes de relacionar partes distantes da imagem desde as camadas iniciais, obtendo um entendimento holístico superior, especialmente quando treinados em grandes volumes de dados.

2.7.2.2.1 Arquitetura PVTv2-B2

O *Pyramid Vision Transformer version 2* (PVTv2), desenvolvido por (WANG *et al.*, 2022), é uma evolução das arquiteturas ViT projetada para superar a limitação de gerar mapas de características em uma única escala. Independentemente da variação escolhida, o PVTv2 introduz uma estrutura piramidal que reduz progressivamente a resolução espacial dos dados enquanto aumenta a profundidade dos canais, assemelhando-se à dinâmica das CNNs clássicas. Isso permite que o modelo extraia desde detalhes finos de alta resolução até características semânticas globais em baixa resolução. Para adequar sua complexidade a diferentes necessidades, o modelo disponibiliza configurações de escalonamento que variam da versão B0 à B5, refletindo o aumento gradual na profundidade da rede e no número de parâmetros, o que permite escolher o melhor compromisso entre custo computacional e capacidade de aprendizado. A Figura 6 ilustra a arquitetura geral do PVTv2.

Figura 6 – Arquitetura PVTv2.



Fonte: (AHMAD *et al.*, 2025)

A variante PVTv2-B2 é gerada a partir da estrutura base, expandindo a capacidade de processamento espacial e os mecanismos de atenção em relação à versão mais enxuta (B0). Enquanto o modelo inicial foi concebido para ser extremamente leve, a versão B2 atua como

uma configuração intermediária que entrega um excelente equilíbrio entre um alto desempenho preditivo e a viabilidade computacional. Devido a essa capacidade aprimorada de capturar dependências de longo alcance e representações globais sem o custo proibitivo das versões B4 e B5, o PVTv2-B2 atua de forma robusta como *backbone* em tarefas complexas de visão computacional (WANG *et al.*, 2022).

2.7.3 Ensemble

O *Ensemble* é uma abordagem consolidada no *Machine Learning* que combina múltiplos modelos preditivos para obter um resultado final mais preciso e robusto. O objetivo principal dessa estratégia é reduzir as fraquezas e os erros individuais de cada classificador, resultando em um sistema com maior capacidade de generalização e menor probabilidade de falhas isoladas (ZHOU, 2012). Entre os métodos mais utilizados para a combinação de modelos, destacam-se o *Ensemble Voting* e o *Ensemble Stacking*.

2.7.3.1 Ensemble Voting

O *Ensemble Voting* é um método de combinação de modelos que define a classificação final a partir dos votos ou da média das probabilidades produzidas pelos modelos base. Essa abordagem pode ser implementada de duas formas principais: *Hard Voting* e *Soft Voting*.

No *Hard Voting*, a classe final corresponde àquela que recebe a maioria dos votos entre os modelos. Por exemplo, em um comitê de três classificadores, se dois predizem uma imagem endoscópica como “anormal” e apenas um como “normal”, a decisão majoritária final será “anormal”. Já no *Soft Voting*, a decisão é obtida pela média das probabilidades previstas por cada modelo, sendo selecionada a classe com maior valor médio (KITTLER *et al.*, 1998). Como ilustração, se as três redes estimam a probabilidade da classe “anormal” em 90%, 80% e 40%, a média resultante de 70% definirá o diagnóstico final como “anormal”, mesmo havendo discordância de um dos modelos quanto à predição absoluta.

O uso dessas abordagens pode contribuir para reduzir a variância do sistema e melhorar a robustez das predições, sem adicionar custos computacionais significativos na etapa de combinação. Entretanto, por utilizarem regras fixas de combinação, não consideram o desempenho individual dos modelos, podendo limitar a capacidade de capturar relações mais complexas entre as predições. Essa limitação pode comprometer o desempenho em cenários com alta variabilidade e ambiguidade visual.

2.7.3.2 Ensemble Stacking

O *Ensemble Stacking* é um método que utiliza as predições de vários modelos base como dados de entrada para treinar um modelo final, adotando uma arquitetura de aprendizado hierárquico. Diferentemente do *Ensemble Voting*, que apenas agrega os resultados dos classificadores, o *Ensemble Stacking* introduz um meta-modelo responsável por aprender a melhor forma de combinar e corrigir as predições geradas pelos modelos iniciais (WOLPERT, 1992a).

Durante esse processo, as saídas dos modelos primários passam a ser utilizadas como novas características (*features*) no treinamento do meta-modelo. Por exemplo, se uma arquitetura convolucional classifica uma imagem endoscópica como "normal" por focar em padrões locais de textura, enquanto um modelo baseado em atenção aprendiz como "anormal" ao capturar o contexto global da lesão, o meta-modelo avalia esse padrão de predições e aprende a atribuir o peso correto a cada rede para emitir o diagnóstico final preciso. Essa estratégia permite que o sistema identifique automaticamente quais classificadores base apresentam maior confiabilidade para determinados padrões de dados, frequentemente superando o desempenho de métodos de combinação mais simples (ZHOU, 2012).

Diversos algoritmos podem ser utilizados como meta-modelo nesse processo, sendo a Regressão Logística uma das abordagens frequentemente empregadas.

2.7.3.2.1 Regressão Logística

A Regressão Logística é um modelo estatístico de *Machine Learning* supervisionado amplamente utilizado para problemas de classificação, cujo objetivo é estimar a probabilidade de uma instância pertencer a uma determinada classe (HASTIE *et al.*, 2009). Para realizar essa tarefa, o algoritmo calcula inicialmente uma combinação linear das variáveis de entrada e seus respectivos pesos. Como esse resultado matemático pode assumir qualquer valor real, aplica-se a função sigmoide, para mapear esse valor em um intervalo contínuo entre 0 e 1. Essa etapa é fundamental, pois converte o cálculo numérico em uma probabilidade interpretável, permitindo a tomada de decisão do classificador. A função sigmoide é definida matematicamente pela Equação 2.1, onde z representa a combinação linear das características de entrada.

$$\text{Sigmoide} = f(z) = \frac{1}{1+e^{-z}}. \tag{2.1}$$

Se a probabilidade obtida superar um limiar predefinido, a amostra é classificada

como positiva, do contrário, como negativa (JR *et al.*, 2013). Pela sua eficiência computacional, robustez contra *overfitting* e saídas probabilísticas, a Regressão Logística atua eficazmente tanto como classificador independente quanto como meta-modelo (meta-classificador) no *Ensemble Stacking*, refinando as predições de redes mais complexas.

2.7.4 Validação Cruzada K-Fold

A Validação Cruzada *K-Fold* (*K-Fold Cross-Validation*) é uma técnica de reamostragem estatística usada para avaliar a capacidade de generalização de modelos de *Machine Learning* e diminuir o risco *overfitting*. A técnica consiste em particionar aleatoriamente o conjunto de dados original em K subconjuntos mutuamente exclusivos e de proporções aproximadamente iguais. Durante o processo de avaliação, o modelo é treinado K vezes. Em cada iteração, um subconjunto diferente é reservado exclusivamente para teste (validação), enquanto os $K-1$ subconjuntos restantes são combinados e empregados na etapa de treinamento (STONE, 1974).

Ao final das K iterações, o desempenho geral do modelo é calculado por meio da média das métricas de avaliação obtidas em cada rodada. Essa abordagem garante que todas as amostras do conjunto de dados sejam utilizadas tanto para treinamento quanto para teste, resultando em uma estimativa de desempenho mais estável e menos dependente de uma única divisão arbitrária dos dados (KOHAVI *et al.*, 1995).

2.8 Métricas de Desempenho

A avaliação do desempenho de modelos de *Machine Learning* é essencial para verificar sua capacidade de generalização e adequação ao problema em estudo. Entre as métricas mais utilizadas em tarefas de classificação estão a acurácia (ACC), sensibilidade ((*recall* - SEN), precisão (PRE), especificidade (ESP) e o F1-score. Cada uma dessas medidas fornece informações complementares, permitindo análises mais equilibradas sobre o comportamento do modelo em diferentes cenários (POWERS, 2011).

Para o cálculo das métricas de desempenho, utiliza-se a matriz de confusão, que considera quatro variáveis fundamentais. O Verdadeiro Positivo (VP) corresponde às instâncias positivas corretamente classificadas pelo modelo. O Falso Positivo (FP) indica instâncias que foram classificadas como positivas incorretamente. O Verdadeiro Negativo (VN) representa as

instâncias negativas corretamente identificadas, enquanto o Falso Negativo (FN) corresponde às instâncias positivas que não foram detectadas pelo modelo, permitindo uma avaliação completa do desempenho do classificador em diferentes cenários (POWERS, 2011; SOKOLOVA; LAPALME, 2009).

A acurácia mede a proporção de previsões corretas em relação ao total de instâncias avaliadas. É uma métrica de fácil interpretação e bastante utilizada como ponto de partida em avaliações experimentais. No entanto, em bases de dados desbalanceadas, a acurácia pode levar a interpretações equivocadas, já que um modelo pode obter alto desempenho apenas favorecendo a classe majoritária, sem de fato aprender padrões relevantes (POWERS, 2011). O cálculo da acurácia é definido na Equação 2.2,

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}. \tag{2.2}$$

A precisão corresponde à proporção de instâncias classificadas corretamente como positivas em relação a todas aquelas que o modelo previu como pertencentes à classe positiva. Essa métrica está diretamente relacionada à minimização de falsos positivos, sendo de especial relevância em aplicações onde previsões incorretas positivas podem trazer consequências críticas, como em diagnósticos médicos ou sistemas de segurança (SOKOLOVA; LAPALME, 2009). O cálculo da precisão é definido na Equação 2.3,

$$\text{Precisão} = \frac{VP}{VP+FP}. \tag{2.3}$$

A sensibilidade, por sua vez, mede a proporção de instâncias positivas corretamente identificadas em relação ao total de elementos realmente pertencentes à classe de interesse. É particularmente importante em contextos nos quais deixar de identificar um caso positivo pode trazer custos significativos, como em sistemas de detecção de doenças, fraudes ou falhas industriais (SOKOLOVA; LAPALME, 2009). O cálculo da sensibilidade é definido na Equação 2.4,

$$\text{Sensibilidade} = \frac{VP}{VP+FN}. \tag{2.4}$$

A especificidade mede a proporção de instâncias negativas corretamente identificadas em relação ao total de elementos que realmente pertencem à classe negativa. Essa métrica é crucial para avaliar a capacidade do modelo em rejeitar falsos positivos, o que, no contexto de diagnósticos médicos, significa reduzir alarmes falsos e evitar intervenções desnecessárias

em pacientes que não apresentam a anomalia (SOKOLOVA; LAPALME, 2009). O cálculo da especificidade é definido na Equação 2.5,

$$\text{Especificidade} = \frac{VN}{VN+FP}. \tag{2.5}$$

O F1-score combina as métricas de precisão e sensibilidade por meio de sua média harmônica, oferecendo uma visão equilibrada do desempenho do modelo. Ele é especialmente útil em cenários de desbalanceamento de classes, nos quais a acurácia sozinha não é capaz de refletir adequadamente a qualidade da classificação. Graças a essa característica, o F1-score tornou-se uma métrica amplamente empregada em problemas reais, servindo como referência para comparar modelos em tarefas complexas de classificação (POWERS, 2011; SOKOLOVA; LAPALME, 2009). O cálculo da F1-score é definido na Equação 2.6,

$$\text{F1-score} = \frac{2VP}{2VP+FP+FN}. \tag{2.6}$$

Todas as métricas apresentadas têm como objetivo fornecer uma análise abrangente do desempenho do modelo, permitindo identificar tanto sua eficácia geral quanto seus pontos de limitação. A utilização conjunta dessas medidas possibilita uma avaliação mais equilibrada, evidenciando aspectos positivos e negativos do classificador. Dessa forma, além de validar o método proposto, os resultados obtidos também servem como subsídio para aprimoramentos em trabalhos futuros.

2.9 Testes de Hipótese

Na avaliação de modelos de *Machine Learning*, a simples comparação das médias das métricas de desempenho pode ser insuficiente para afirmar que um classificador é estritamente superior a outro, uma vez que a diferença observada pode ser fruto de variações aleatórias na divisão dos dados de treinamento e teste. Para garantir o rigor analítico, empregam-se testes de hipótese estatísticos, que avaliam a probabilidade de as diferenças de desempenho serem significativas (DEMŠAR, 2006). Nesses testes, a hipótese nula (H_0) assume que não há diferença real entre os algoritmos avaliados, enquanto a hipótese alternativa (H_1) sugere que a divergência observada é estatisticamente válida.

2.9.1 Teste t-pareado

O teste t-pareado (*Paired t-test*) é um método estatístico paramétrico utilizado para comparar as médias de duas amostras dependentes. No contexto de *Machine Learning*, é frequen-

temente aplicado para comparar o desempenho de dois classificadores testados sobre os mesmos agrupamentos de dados (*folds*) gerados pela validação cruzada. O método calcula a diferença entre os resultados pareados e avalia se a média dessas diferenças difere significativamente de zero. Para que o teste mantenha sua validade teórica, assume-se que as diferenças entre os pares sigam uma distribuição normal. No entanto, essa suposição de normalidade, assim como a independência das amostras oriundas de reamostragem, muitas vezes é violada em experimentos com classificadores, o que pode comprometer a confiabilidade do teste paramétrico (DEMŠAR, 2006).

2.9.2 Teste de Wilcoxon

Para contornar as limitações do método paramétrico, utiliza-se o teste dos postos sinalizados de Wilcoxon (*Wilcoxon signed-rank test*), que atua como a alternativa não-paramétrica ao teste t-pareado. Esse teste não exige que as diferenças sigam uma distribuição normal, sendo considerado uma abordagem mais segura e robusta para a comparação de algoritmos (DEMŠAR, 2006). Em vez de utilizar as médias, o método calcula as diferenças absolutas de desempenho entre os pares, ordena esses valores atribuindo-lhes postos (*ranks*) e, em seguida, soma os postos correspondentes às diferenças positivas e negativas. Caso a divergência de postos seja acentuada a favor de um dos modelos, a hipótese nula é rejeitada, indicando a existência de diferença estatisticamente significativa entre as abordagens avaliadas, sem a necessidade de assumir premissas rígidas sobre a distribuição dos dados.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta diversos estudos que utilizam *Deep Learning* para a classificação multiclasse de doenças gastrointestinais em imagens endoscópicas.

No contexto da classificação de doenças gastrointestinais, (AYAN, 2024) propuseram uma abordagem comparativa de *Deep Learning* utilizando *Vision Transformers* e CNNs. O modelo de maior destaque emprega a arquitetura DenseNet201, aprimorada com parâmetros otimizados de *Transfer Learning* para a extração de características em imagens endoscópicas. O método alcançou acurácia de 93,13% e F1-score de 93,11%. Aprofundando a integração entre essas arquiteturas, (SUBEDI *et al.*, 2024) desenvolveram um modelo híbrido que também utiliza a rede DenseNet201 para a extração de características locais, mas a integra ao *Swin Transformer* para a compreensão global do contexto visual. Essa combinação visa aprimorar a robustez do diagnóstico em cenários com classes desbalanceadas. O método obteve acurácia de 72,39% e F1-score de 69%.

No trabalho (DEMIRBA *et al.*, 2024) foi desenvolvida uma arquitetura híbrida multiclasse *Spatial-Attention ConvMixer*, que combina o *ConvMixer* com um mecanismo de atenção espacial. A abordagem alcançou acurácia de 93,37% e F1-score de 93,42%. Buscando reduzir o problema de *overfitting* comum em imagens médicas complexas, (SIDDIQUI *et al.*, 2025) apresentaram uma abordagem baseada em um modelo *Deep Ensemble Network* customizado. O trabalho incorpora técnicas de *Transfer Learning* para extrair características visuais profundas e, em avaliação cruzada, atingiu acurácia de 97,80% e F1-score de 96%.

De forma semelhante, (SEHMUS, 2025) apresentaram um *framework* de *Stacking Ensemble* em dois níveis para a classificação de imagens endoscópicas. No primeiro nível, três arquiteturas de redes neurais pré-treinadas ResNet50, DenseNet201 e MobileNetV3Large foram utilizadas simultaneamente como modelos base para extrair características e gerar previsões. No segundo nível, as previsões dessas três redes foram combinadas e passaram a servir como entrada para um meta-classificador. Ao avaliar diferentes algoritmos clássicos para atuar nesse segundo nível, o modelo atingiu sua melhor performance utilizando o *Random Forest* como o meta-classificador final. Essa arquitetura obteve uma acurácia de 94,33% e um F1-score de 94,27% , comprovando que a união de múltiplos extratores reduz as limitações das redes individuais na distinção de classes visualmente semelhantes.

Embora esses métodos multiclasse sejam eficazes, observa-se uma escassez de abordagens voltadas para a classificação binária na literatura. A distinção rápida entre exames

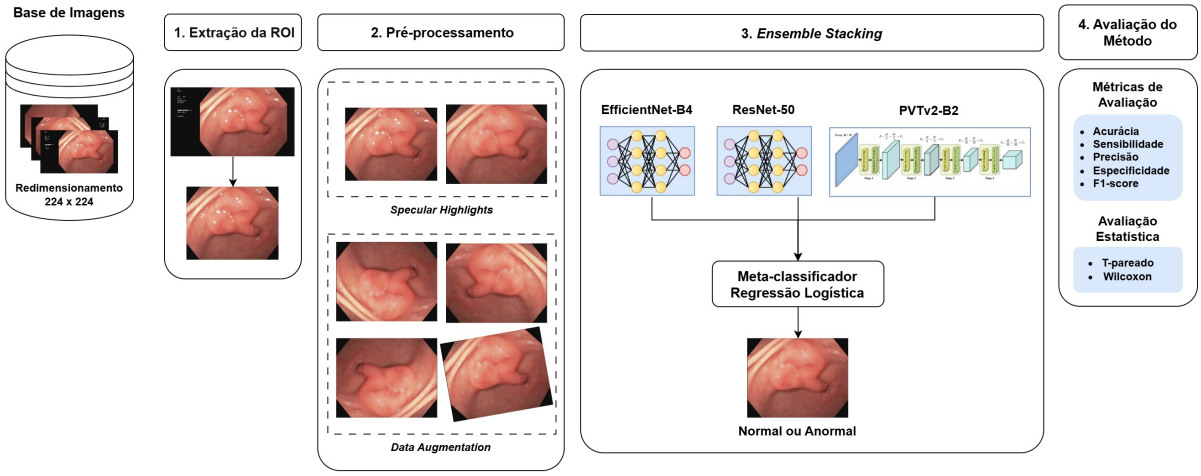
normais e anormais é particularmente importante em cenários de triagem, pois permite priorizar casos suspeitos, favorecendo a intervenção precoce e a alocação mais eficiente de recursos no sistema de saúde.

Além disso, a literatura explora diferentes estratégias para aprimorar a análise de imagens endoscópicas, como arquiteturas profundas para extração de características, modelos híbridos que combinam CNNs e *Transformers*, e técnicas de *Ensemble*. Embora essas abordagens apresentem resultados promissores, elas são frequentemente empregadas de forma isolada, limitando o potencial entre diferentes extratores de características. Nesse contexto, este trabalho propõe um método que integra essas estratégias e incorpora etapas de pré-processamento, como extração da ROI, redução de SH e DA, juntamente com uma técnica baseada em *Ensemble Stacking* que combina CNNs e *Vision Transformers* para dar suporte à triagem automática de exames endoscópicos.

4 MATERIAIS E MÉTODO PROPOSTO

O método proposto compreende quatro etapas, validadas utilizando a base de imagens apresentada na subseção 4.1. Na primeira etapa, foram removidas as áreas irrelevantes por meio da extração da ROI. Na segunda etapa, foi realizado o pré-processamento, que incluiu a redução de reflexos de SH e a aplicação de técnicas de DA para diversificar a base. Na terceira etapa, diversas arquiteturas de CNNs e ViT foram treinadas e combinadas usando *Ensemble Stacking* para a construção do modelo preditivo. Por fim, na quarta etapa, foi realizada a avaliação do método por meio das métricas de avaliação e da análise estatística, permitindo mensurar a robustez dos resultados. A Figura 7 ilustra o processo descrito. A Figura 7 ilustra o processo descrito.

Figura 7 – Ilustração do método proposto.



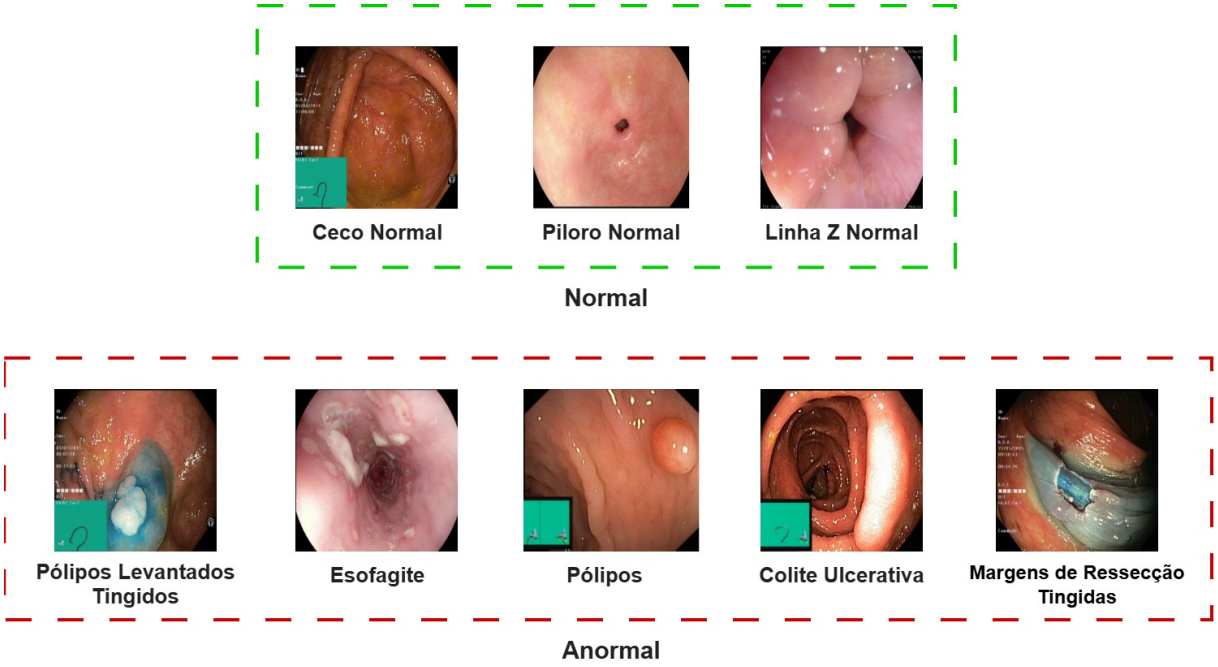
Fonte: Elaborado pelo autor.

4.1 Base de Imagens

Os dados foram coletados da base de imagens público *Kvasir V1* no *Kaggle* (POGORELOV *et al.*, 2017), composto por 4000 imagens endoscópicas em oito classes: Pólipos Levantados com Corante (*Dyed Lifted Polyps* – DLP), (*Normal Cecum* – NC), Piloro Normal (*Normal Pylorus* – NP), Linha Z Normal (*Normal Z Line* – NZL), Margens de Ressecção Tingidas (*Dyed Resection Margins* – DRM), Esofagite (*Esophagitis* – ESO), Pólipos (*Polyps* – P) e Colite Ulcerativa (*Ulcerative Colitis* – UCE), com 500 imagens cada. As dimensões variam de 730x576 a 1920x1072 *pixels*, todas em formato JPG. O método proposto distingue entre classes normais e anormais: NC, NP e NZL (1500 imagens) são normais, enquanto DLP, DRM,

ESO, P e UCE (2500 imagens) são anormais. Essa base foi escolhida devido ao seu amplo reconhecimento na literatura e ao fato de ter sido construída e anotada por especialistas médicos. A Figura 8 apresenta exemplos das classes utilizadas neste estudo.

Figura 8 – Separação das classes da base nas categorias normal e anormal.



Fonte: Elaborado pelo autor.

Conforme demonstrado por (THAMBAWITA *et al.*, 2021), resoluções mais elevadas maximizam o desempenho preditivo em imagens endoscópicas, porém implicam aumento significativo no custo computacional e no consumo de memória de vídeo (GPU). Considerando que o método proposto utiliza uma abordagem de *Ensemble Stacking*, que requer o treinamento e a inferência de múltiplos modelos de alta complexidade, o uso de altas resoluções tornaria inviável em cenários com recursos computacionais limitados. Dessa forma, neste estudo, as imagens foram padronizadas para 224×224 *pixels*, valor definido por representar um equilíbrio adequado entre custo computacional e desempenho, preservando as características morfológicas essenciais das lesões e garantindo a viabilidade do método.

4.2 Extração da ROI

Algumas imagens contêm áreas irrelevantes, como bordas pretas usadas para anotações médicas, que não contribuem para o aprendizado e podem enviesar o modelo. Para minimizar este problema, foi aplicada a extração de ROI (VIANA *et al.*, 2024), que consiste em

selecionar o maior quadrado possível que não contenha pixels pretos em suas laterais centrais, formando uma caixa delimitadora que preserva a região central útil da imagem. Dessa forma, as áreas periféricas irrelevantes são removidas, mantendo-se apenas as regiões mais informativas para o diagnóstico (Figura 9).

Figura 9 – Extração da ROI.

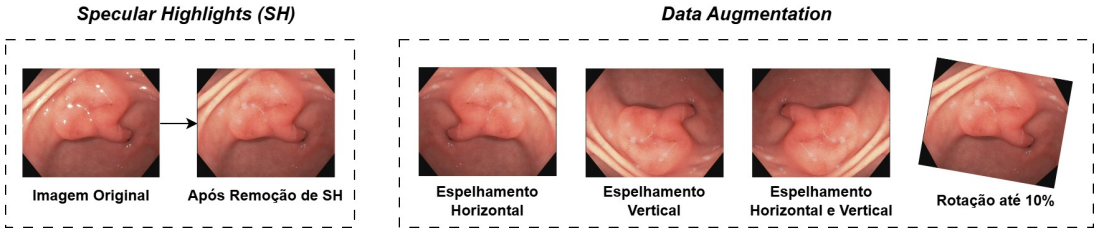
Fonte: Elaborado pelo autor.

4.2.1 Pré-processamento

Esta etapa é composta por dois processos complementares: a redução de *Specular Highlights* (SH) e o *Data Augmentation* (DA).

No primeiro, a incidência da iluminação endoscópica sobre a mucosa úmida frequentemente gera reflexos especulares que ocluem texturas relevantes e introduzem ruído visual. Esses artefatos podem degradar informações importantes do tecido e comprometer a extração de características discriminativas pelos modelos de *Deep Learning*. Para reduzir a interferência desses artefatos, realiza-se a segmentação dos *pixels* cuja luminância excede um limiar pré-definido. Em seguida, aplica-se a técnica de *inpainting*, que reconstrói as áreas corrompidas interpolando informações da vizinhança adjacente, restaurando assim a integridade morfológica do tecido. No segundo processo, aplica-se DA com rotações aleatórias de até 10% e espelhamentos horizontal e/ou vertical (Figura 10). Esta estratégia diversifica a base de treinamento gerando novas variações das imagens originais, favorecendo a capacidade de generalização dos modelos e contribuindo para a redução do *overfitting*.

Figura 10 – Processos da etapa de pré-processamento.



Fonte: Elaborado pelo autor.

4.2.2 Ensemble Stacking

A etapa de classificação utiliza o *Ensemble Stacking* (WOLPERT, 1992b) para extrair o máximo potencial da diversidade entre CNNs e ViT. Neste contexto, a meta-classificação é realizada pelo algoritmo de Regressão Logística (RL) (JR *et al.*, 2013), que mapeia o espaço de predições dos modelos base para a decisão final. Ao invés de uma agregação estática, como *Soft Voting* (AWE *et al.*, 2024), a RL atua como um mecanismo de ponderação dinâmica, discernindo padrões de erro específicos e atribuindo pesos maiores à rede que apresenta maior competência para cada amostra da base de imagens, evitando o problema de seleção arbitrária de modelos.

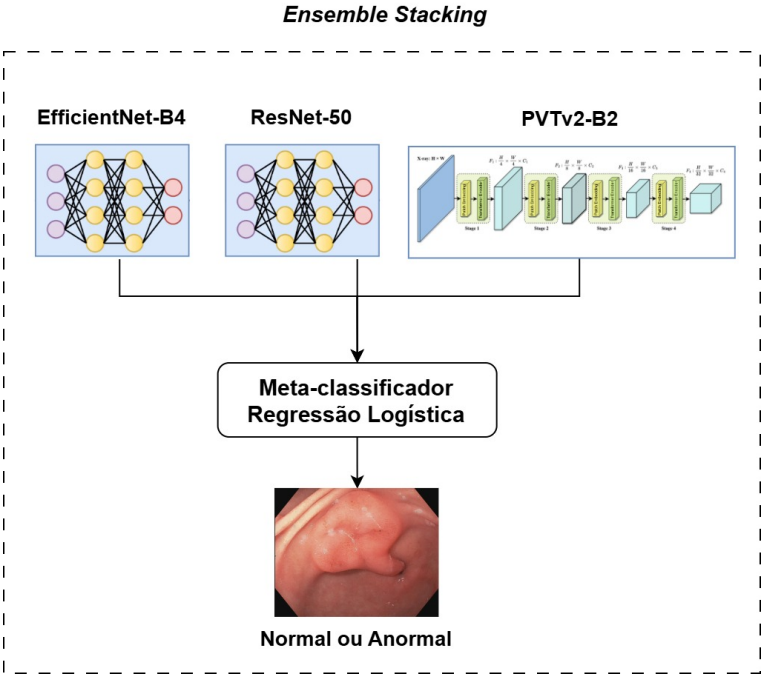
Foram selecionadas arquiteturas complementares para compor o *Ensemble Stacking*, incluindo EfficientNet-B4, ResNet-50 e o PVTv2-B2. Esse conjunto combina CNNs com diferentes capacidades de extração de características e uma arquitetura baseada em ViT, permitindo explorar diferentes representações dos padrões visuais. Essa combinação favorece a complementaridade entre os modelos, reduzindo o viés de arquiteturas isoladas e aumentando a robustez na classificação de imagens endoscópicas (HUSSAIN *et al.*, 2025). Os modelos foram treinados individualmente e suas saídas foram concatenadas no meta-classificador para uma decisão unificada (pode ser visualizada na Figura 11).

Para avaliar o desempenho do método proposto, foram empregadas métricas de avaliação e análise estatística, permitindo mensurar a robustez e a consistência dos resultados obtidos.

4.2.2.1 Métricas de Avaliação

A eficácia do método preditivo final foi avaliada por meio de métricas padrão da literatura (Seção 2.8): acurácia (ACC), sensibilidade (SEN), precisão (PRE), especificidade (ESP) e F1-score (DUDA, 1973).

Figura 11 – *Ensemble Stacking*.



Fonte: Elaborado pelo autor.

4.2.2.2 Avaliação Estatística

Para a avaliação estatística, foram utilizados os testes de hipótese t-pareado e *Wilcoxon signed-rank* (Seção 2.9) para verificar se as diferenças de desempenho entre os experimentos avaliados são estatisticamente significativas.

5 RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os experimentos realizados para avaliar o impacto de cada técnica do método proposto. Para isso, foi conduzida uma série de experimentos sistemáticos por meio de um teste de ablação, considerando etapas como extração da ROI, técnicas de pré-processamento (SH e DA), avaliação das CNNs e *Transformers* e comparação entre diferentes métodos de *Ensemble*. Além disso, são apresentados estudos de caso qualitativos e uma comparação com trabalhos da literatura.

5.1 Configuração Experimental

O treinamento dos modelos foi realizado em um ambiente virtual do *Google Colab*, utilizando um *Jupyter Notebook*. As especificações de *hardware* incluíram GPU Nvidia A100 e GPU Nvidia T4 Tensor Core. O desenvolvimento foi conduzido em linguagem Python, usando principalmente as bibliotecas OpenCV, Pandas, Skicit-Learn e Keras.

Os modelos baseados em CNNs e *Transformers* foram treinados utilizando a técnica de *transfer learning*, com pesos pré-treinados na base ImageNet. Para garantir consistência experimental, todas as arquiteturas foram treinadas utilizando os mesmos hiperparâmetros: otimizador *Adam*, taxa de aprendizado inicial de 0.0005, *batch size* de 16, camada de *dropout* de 0,5 na etapa final, função de perda *Categorical Crossentropy*. O treinamento foi conduzido por até 50 épocas, utilizando um critério de *early stopping* para evitar *overfitting*.

Nos testes de ablação, a base de imagens foi dividida aleatoriamente em 70% para treinamento, 10% para validação e 20% para teste. Para a avaliação estatística do método de *Ensemble Stacking*, foi utilizada validação cruzada com $k = 10$ *folds*. A partir dos resultados obtidos em cada *fold*, foram aplicados os testes de hipótese t-pareado e *Wilcoxon signed-rank* (Seção 2.9), a fim de verificar se as diferenças de desempenho entre os métodos avaliados são estatisticamente significativas.

5.1.1 Experimento com e sem Extração da ROI

Para avaliar o impacto da extração da ROI no desempenho da classificação, foram conduzidos experimentos utilizando a arquitetura EfficientNet-B4 como modelo base. Inicialmente, o modelo foi treinado utilizando as imagens originais, sem qualquer etapa de pré-processamento. Em seguida, realizou-se um segundo experimento no qual foi aplicada a

técnica de extração da ROI. A Tabela 1 apresenta os resultados obtidos em cada cenário.

Tabela 1 – Resultados do experimento com e sem a extração da ROI.

Experimento	**ACC**	**PRE**	**SEN**	**ESP**	**F1-score**
Imagens Originais	91,50%	91,48%	91,50%	90,80%	91,49%
Extração da ROI	**92,25%**	**92,26%**	**92,25%**	**91,80%**	**92,26%**

Fonte: Elaborado pelo autor.

s resultados sugerem melhorias em todas as métricas ao empregar a extração da ROI. Uma possível explicação é que as bordas irrelevantes da endoscopia continham informações ruidosas que poderiam confundir o classificador. Ao eliminar essas áreas desnecessárias, o modelo alcançou uma capacidade de generalização superior, resultando em maior eficiência e precisão na classificação das imagens endoscópicas.

5.1.2 Experimento com e sem Pré-processamento

Na Seção 4.2.1, discute-se a etapa de pré-processamento utilizada neste estudo, composta de forma conjunta pela redução de SH e pela técnica de DA. Para validar a abordagem proposta, foram conduzidos dois experimentos complementares: o primeiro utilizou as imagens apenas com a extração da ROI, enquanto o segundo incorporou as técnicas de pré-processamento (SH e DA). Os experimentos foram realizados mantendo o modelo EfficientNet-B4. Os resultados estão sumarizados na Tabela 2.

Tabela 2 – Resultados do experimento com e sem a etapa de pré-processamento.

Experimento	**ACC**	**PRE**	**SEN**	**ESP**	**F1-score**
Extração da ROI	92,25%	92,26%	92,25%	91,80%	92,26%
Extração da ROI + DA + SH	**93,37%**	**93,66%**	**93,37%**	**93,42%**	**93,77%**

Fonte: Elaborado pelo autor.

Conforme evidenciado pelos resultados na Tabela 2, a implementação do pré-processamento resultou em melhorias em todas as métricas avaliadas. Esse ganho pode ser atribuído, em parte, à redução de SH, cuja presença pode degradar informações visuais relevantes e comprometer a extração de características discriminativas. A atenuação desses artefatos contribui para preservar padrões estruturais do tecido e favorece representações mais discriminativas para o processo de classificação. Aliada a isso, a aplicação de DA diversifica a base de treinamento, favorecendo a capacidade de generalização e redução de *overfitting*, resultando em classificações mais robustas pelo método proposto.

5.1.3 Experimento com Modelos Individuais

Nesta seção, foram avaliadas sete arquiteturas amplamente utilizadas na literatura, incluindo modelos baseados em CNNs e *Transformers*, a fim de estabelecer um desempenho base para a tarefa de classificação de anomalias em imagens endoscópicas. Ressalta-se que todos os modelos foram treinados utilizando extração da ROI e etapas de pré-processamento. A Tabela 3 apresenta os resultados obtidos, organizados em ordem de melhor F1-score.

Tabela 3 – Resultados dos modelos individuais com extração da ROI e pré-processamento.

Modelo	**ACC**	**PRE**	**SEN**	**ESP**	**F1-score**
PVTv2-B2	96,37%	96,48%	96,37%	96,63%	96,39%
ResNet-50	93,87%	94,45%	93,87%	94,30%	93,92%
EfficientNet-B4	93,37%	93,66%	93,37%	93,42%	93,77%
ViT-32	93,50%	93,97%	93,50%	94,20%	93,56%
VGG-19	92,63%	92,71%	92,63%	92,50%	92,65%
MobileNetV2	92,37%	92,37%	92,37%	91,83%	92,37%
InceptionV3	92,37%	92,36%	92,37%	91,43%	92,34%

Fonte: Elaborado pelo autor.

A partir dos resultados apresentados na Tabela 3, observa-se que a arquitetura PVTv2-B2 apresentou o melhor desempenho geral, com F1-score de 96,39%. Esse resultado evidencia a capacidade dos modelos baseados em *Transformers* de capturar dependências globais nas imagens por meio de mecanismos de atenção. Por outro lado, o ViT-32, também baseado nessa abordagem, apresentou desempenho inferior, com F1-score de 93,56%, possivelmente devido à maior dependência desse tipo de arquitetura por grandes volumes de dados para treinamento.

De modo geral, arquiteturas baseadas em CNNs também apresentaram desempenho competitivo. As CNNs exploram principalmente padrões espaciais locais por meio de operações convolucionais, o que é particularmente relevante em tarefas de análise de imagens endoscópicas, nas quais texturas e padrões visuais locais podem indicar a presença de anomalias. Nesse contexto, a ResNet-50 destacou-se como a melhor CNN avaliada, obtendo F1-score de 93,92%, seguida pela EfficientNet-B4 com 93,77%. Esses resultados demonstram a eficiência de arquiteturas profundas e com mecanismos de otimização estrutural para extração de características relevantes.

Considerando o desempenho obtido, os três modelos com maiores valores de F1-score (PVTv2-B2, ResNet-50 e EfficientNet-B4) foram selecionados para compor o método de *Ensemble Stacking*. A escolha desses modelos busca explorar a diversidade entre arquiteturas baseadas em *Transformers* e CNNs, combinando suas diferentes capacidades de extração de

características para obter um modelo final mais robusto e preciso.

5.1.4 Comparação entre Métodos de Ensemble

Após a seleção das três arquiteturas com melhor desempenho e visando explorar a complementaridade entre suas representações, foram avaliadas três estratégias de combinação: *Soft Voting*, *Hard Voting* e *Ensemble Stacking*. A Tabela 4 apresenta os resultados comparativos dessas abordagens.

Tabela 4 – Resultados comparativos das diferentes técnicas de *Ensemble*.

Experimento	ACC	PRE	SEN	ESP	F1-score
Ensemble Soft Voting	96,88%	97,05%	96,88%	97,37%	96,89%
Ensemble Hard Voting	96,88%	97,08%	96,88%	97,43%	96,90%
Ensemble Stacking	**98,12%**	**98,15%**	**98,12%**	**98,23%**	**98,13%**

Fonte: Elaborado pelo autor.

Os resultados apresentados na Tabela 4 indicam que a estratégia *Ensemble Stacking* obteve o melhor desempenho entre as abordagens avaliadas. Enquanto os métodos *Soft Voting* e *Hard Voting* combinam as predições dos modelos por meio de regras de combinação fixas, usando probabilidades médias e votação majoritária, respectivamente, o *Ensemble Stacking* emprega um meta-classificador baseado em RL para aprender a integrar as saídas dos modelos base. Dessa forma, explora a complementaridade entre CNNs e ViT, resultando em uma combinação mais eficaz de predições e melhorias consistentes nas métricas de desempenho.

5.1.5 Evolução do Método Proposto

Finalmente, foi conduzido um estudo de ablação utilizando a arquitetura EfficientNet-B4 como modelo base para demonstrar a evolução do método proposto e quantificar a contribuição de cada etapa. O experimento incorporou progressivamente a extração do ROI, a etapa de pré-processamento e, por fim, a substituição do modelo individual pelo método final de *Ensemble Stacking*. Os resultados dessa análise estão apresentados na Tabela 5.

O estudo de ablação destaca a contribuição progressiva de cada etapa do método proposto. A inclusão da etapa de extração da ROI promoveu o primeiro ganho de desempenho, indicando que a remoção de bordas e anotações irrelevantes reduz ruídos que podem prejudicar o aprendizado do modelo. Em seguida, a etapa de pré-processamento, composta pela redução de SH e pela aplicação de DA, resultou em novos avanços, refletindo uma melhoria na capacidade

Tabela 5 – Resultados do estudo de ablação para as diferentes configurações do método proposto.

Experimento	ACC	PRE	SEN	ESP	F1-score
EfficientNet-B4	91,50%	91,48%	91,50%	90,80%	91,49%
EfficientNet-B4 + ROI	92,25%	92,26%	92,25%	91,80%	92,26%
EfficientNet-B4 + ROI + SH + DA	93,37%	93,66%	93,37%	93,42%	93,77%
***Ensemble Stacking* + ROI + SH + DA**	**98,12%**	**98,15%**	**98,12%**	**98,23%**	**98,13%**

Fonte: Elaborado pelo autor.

de generalização do modelo individual. O melhor desempenho é obtido com a adição do *Ensemble Stacking*, que combina as predições dos modelos complementares. Comparado aos resultados iniciais, observou-se um aumento de 6,62% na acurácia, 6,67% na precisão, 6,62% na sensibilidade, 7,43% na especificidade e 6,64% no F1-score, demonstrando a contribuição combinada das etapas propostas para a classificação de imagens endoscópicas.

5.1.6 Análise Estatística dos Métodos de Ensemble

Neste experimento foi realizada uma análise estatística para avaliar se as melhorias obtidas pelo método proposto são estatisticamente significativas, comparando duas configurações do método *Ensemble Stacking*: uma utilizando as imagens originais, sem extração de ROI, sem aplicação de SH e sem DA, e outra utilizando o método proposto completo. Essa análise foi conduzida por se tratar da etapa final e mais relevante da abordagem proposta.

Para isso, foi utilizada validação cruzada com $k = 10$ *folds*, a partir dos quais foram calculadas as médias e os desvios padrão das métricas. Além disso, foram aplicados dois testes de hipótese para verificar a significância estatística das diferenças entre os métodos (Seção 2.9): o teste t-pareado e o teste não paramétrico de *Wilcoxon signed-rank*. Considerando um nível de significância de 5% ($\alpha = 0.05$), valores de p inferiores a esse limiar indicam que as diferenças observadas entre os métodos são estatisticamente significativas. Os resultados consolidados dessa comparação, juntamente com os p-valores obtidos, estão dispostos na Tabela 6.

Tabela 6 – Comparação de desempenho entre os métodos e resultados dos testes de significância estatística.

Método/Testes de Hipóteses	ACC(%)	PRE(%)	SEN(%)	ESP(%)	F1-score(%)
Ensemble Stacking	94,80 ± 1,05	94,85 ± 1,04	94,80 ± 1,05	94,53 ± 1,16	94,80 ± 1,05
Ensemble Stacking + ROI + SH + DA	**96,20 ± 0,87**	**96,27 ± 0,86**	**96,20 ± 0,87**	**96,16 ± 0,80**	**96,21 ± 0,86**
Teste T-pareado (p-valor)	0,010	0,009	0,010	0,020	0,011
Teste de *Wilcoxon* (p-valor)	0,039	0,020	0,039	0,020	0,039

Fonte: Elaborado pelo autor.

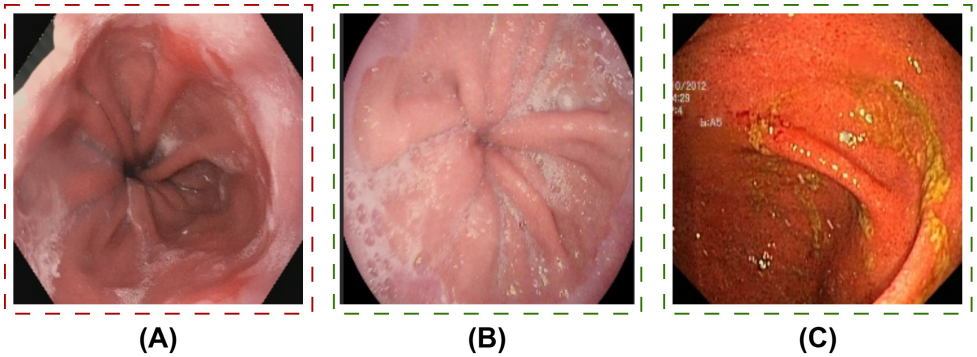
Conforme observado nos resultados, a abordagem proposta obteve métricas mais

altas que o método base (*Ensemble Stacking* isolado), elevando a acurácia média de 94,80% para 96,20% e apresentando desvios padrão menores, o que sugere uma possível maior estabilidade e confiabilidade nas predições dos modelos. Em relação à análise estatística, os testes (t-pareado e Wilcoxon) retornaram *p*-valores estritamente inferiores ao limiar de 0,05 para todas as métricas. Pode-se interpretar que dado o *p*-valor obtido, ao assumir a hipótese nula como verdadeira, há apenas 1% de chance de observar os valores apresentados (ou valores mais extremos). Dada essa baixa probabilidade, rejeita-se a hipótese nula. Logo, os dados sugerem que existe diferença estatisticamente significativa entre os desempenhos do *Ensemble Stacking* isolado e com a integração das etapas (ROI, SH e DA), indicando que essa abordagem possivelmente contribui para a melhoria do diagnóstico das anomalias.

5.1.7 Estudos de Caso

Esta seção apresenta três estudos de caso qualitativos envolvendo a análise de classificações realizadas pelo método proposto, ilustrados na Figura 12. O primeiro caso envolve uma imagem normal erroneamente classificada como anormal (A), caracterizando um falso positivo. O segundo caso apresenta uma imagem anormal corretamente classificada como anormal (B), representando um verdadeiro positivo. Finalmente, o terceiro caso corresponde a uma imagem normal corretamente classificada como normal, caracterizando um verdadeiro negativo (C).

Figura 12 – Estudos de caso. (A) Imagem normal classificada erroneamente como anormal; (B) Imagem anormal classificada corretamente como anormal; e (C) Imagem normal classificada corretamente como normal.



Fonte: Elaborado pelo autor.

Ao analisar a Figura 12 (A), que exibe uma imagem normal, observa-se que as pregas gástricas ou esofágicas são muito proeminentes e apresentam uma coloração avermelhada intensa. O método pode ter interpretado a textura da vascularização do tecido e essa vermelhidão

natural como sinais clínicos de inflamação, como a esofagite, resultando em um diagnóstico de falso positivo.

Em contraste, ao analisar a Figura 12 (B), que exibe uma imagem normal classificada corretamente pelo método. Observa-se claramente a presença de bolhas e fluidos gástricos na superfície do tecido e nas fendas anatômicas. Esses elementos representam um ruído visual significativo, pois distorcem a textura real da mucosa e podem criar falsos padrões que frequentemente confundem os processos tradicionais de extração de características. O acerto na predição demonstra que o método foi robusto o suficiente para ignorar esses artefatos inerentes ao exame e reconhecer os padrões globais do tecido saudável, evitando um falso positivo.

De forma semelhante, a Figura 12 (C) apresenta um caso de Colite Ulcerativa classificado corretamente como anormal, apesar da ambiguidade visual causada por substâncias amareladas que podem corresponder tanto a exsudato inflamatório quanto a resíduos fisiológicos. A predição correta sugere que o método conseguiu capturar relações espaciais entre essas estruturas e o tecido adjacente, superando limitações de abordagens baseadas apenas em cor ou textura local.

Apesar da ocorrência de classificações incorretas em alguns casos ambíguos, os resultados quantitativos indicam que o modelo apresenta alta robustez geral. Quando integrado à expertise médica, o método pode desempenhar um papel crucial como ferramenta de segunda opinião, reduzindo erros humanos e auxiliando na detecção precoce de anormalidades gastrointestinais.

5.1.8 Comparação com a Literatura

Esta seção propõe uma análise comparativa com os estudos discutidos no Capítulo 3. É importante ressaltar que as metodologias empregadas e a quantidade de classes avaliadas diferem significativamente entre os trabalhos, fazendo com que essa comparação não seja absoluta, mas sim uma tentativa de estabelecer um parâmetro de desempenho frente ao estado da arte. A Tabela 7 resume os resultados alcançados na comparação.

Diversos estudos na área empregam arquiteturas profundas para detectar anormalidades em exames endoscópicos. Trabalhos como (AYAN, 2024) e (DEMIRBA *et al.*, 2024) exploraram arquiteturas robustas de extração de características, como a DenseNet201 e o *ConvMixer* com atenção espacial, alcançando valores de F1-score próximos de 93%. Embora eficazes, arquiteturas individuais podem apresentar limitações em capturar simultaneamente características

Tabela 7 – Comparação de trabalhos relacionados e do método proposto.

Trabalho	Método	ACC	PRE	SEN	ESP	F1-score
(AYAN, 2024)	DenseNet201 (*Transfer Learning*)	93,13%	93,13%	93,17%	-	93,11%
(SUBEDI *et al.*, 2024)	DenseNet201 + *Swin Transformer*	72,39%	70,07%	72,39%	-	69%
(DEMIRBA *et al.*, 2024)	*Spatial-Attention ConvMixer*	93,37%	93,66%	93,37%	-	93,42%
(SIDDIQUI *et al.*, 2025)	*Deep Ensemble Learning*	97,80%	98%	97%	-	96%
(SEHMUS, 2025)	*Ensemble Stacking*	94,33%	94,36%	94,27%	-	94,27%
Método Proposto	***Ensemble Stacking* + ROI + SH + DA**	**98,12%**	**98,15%**	**98,12%**	**98,23%**	**98,13%**

Fonte: Elaborado pelo autor.

locais e o contexto global da mucosa.

Nesse cenário, abordagens que integram múltiplos modelos têm sido investigadas para explorar diferentes representações visuais. O trabalho de (SUBEDI *et al.*, 2024), por exemplo, propôs um modelo híbrido combinando CNNs e *Transformers*. No entanto, a ausência de etapas mais elaboradas de pré-processamento resultou em desempenho inferior (F1-score de 69%). De forma semelhante, (SEHMUS, 2025) utilizaram *Ensemble Stacking* para combinar extratores convolucionais, alcançando F1-score de 94,27%, enquanto (SIDDIQUI *et al.*, 2025) aplicaram técnicas de *Deep Ensemble Learning*, obtendo F1-score de 96%.

O método proposto neste estudo avança em relação a essas abordagens ao integrar arquiteturas de naturezas distintas por meio de *Ensemble Stacking*. Diferentemente de estratégias baseadas em agregações fixas, como em (SIDDIQUI *et al.*, 2025), ou na combinação de modelos de mesma natureza predominantemente convolucional, como em (SEHMUS, 2025), o meta-classificador usado aprende como ponderar as saídas das redes base, explorando a complementaridade entre CNNs e ViT. Essa estratégia permite combinar representações locais e globais de forma mais eficaz, favorecendo a identificação de padrões visuais complexos presentes em imagens endoscópicas.

Em comparação com os trabalhos analisados, o método proposto apresentou desempenho superior, alcançando 98,12% de acurácia e 98,13% de F1-score. Esses resultados superam o melhor desempenho relatado na literatura comparativa, obtido por (SIDDIQUI *et al.*, 2025), que alcançou 97,80% de acurácia e F1-score de 96%. Esses resultados demonstram a eficácia da integração entre CNNs e ViT por meio de *Ensemble Stacking*, aliada às etapas como extração de ROI, redução de SH e DA. No contexto clínico, a alta acurácia indica maior confiabilidade na classificação geral dos exames, enquanto o alto F1-score reflete um melhor equilíbrio entre precisão e sensibilidade, contribuindo para reduzir a ocorrência de falsos negativos, o que é de suma importância para evitar que anomalias precursoras de câncer passem despercebidas. Dessa forma, o método proposto apresenta potencial para apoiar a triagem e promover intervenções

clínicas mais seguras e precoces.

5.2 Aspectos Importantes do Método Proposto

A classificação de anomalias gastrointestinais em imagens endoscópicas não é uma tarefa trivial devido à presença de artefatos visuais e à similaridade estrutural entre tecidos normais e patológicos. Desenvolver um método capaz de contornar essas adversidades e atingir uma alta taxa de acerto é um desafio significativo. Neste trabalho, as principais etapas do método proposto mostraram-se robustas e precisas. A partir da análise dessas etapas, destacam-se os principais aspectos e limitações observados, os quais são discutidos a seguir.

1. O estudo em questão apresenta um método completo para resolver o problema de triagem e classificação binária de exames endoscópicos. Nesta área, a complexidade visual é amplamente reconhecida, e os resultados obtidos ganham lugar de destaque frente aos métodos de última geração da literatura, alcançando 98,12% de acurácia e 98,13% de F1-score.
2. A aplicação da extração da ROI foi um passo fundamental. Ao remover bordas pretas e anotações médicas irrelevantes, eliminou-se uma fonte de informação ruidosa, o que permitiu que os modelos alcançassem uma capacidade de generalização superior e focassem apenas nas áreas clinicamente úteis.
3. Em relação ao pré-processamento, a união da redução de SH com o DA causou um impacto positivo. A atenuação de reflexos de iluminação endoscópica através de *inpainting* restaurou a integridade morfológica do tecido. Simultaneamente, o DA atuou reduzindo o *overfitting*, garantindo que os modelos aprendessem padrões mais robustos e generalizáveis.
4. Um aspecto crucial para o alto desempenho foi a combinação de arquiteturas baseadas em CNNs (EfficientNet-B4 e ResNet-50), responsáveis pela extração de características locais, com ViT (PVTv2-B2), que contribuem para a modelagem de relações globais na imagem, resultando em uma representação mais completa e eficaz do tecido analisado.
5. O uso do *Ensemble Stacking* guiado por um meta-classificador de RL foi o fator definitivo para a precisão final. Em vez de aplicar uma regra estática de agregação (como no *Soft Voting*), o modelo aprendeu a ponderar dinamicamente as predições, identificando padrões de erro específicos e priorizando a arquitetura com maior competência para cada amostra.

Embora o método proposto apresente vários fatores positivos e alta precisão, ele

também possui algumas limitações, nas quais destacam-se:

1. O método proposto demonstrou suscetibilidade a gerar falsos positivos diante de ambiguidades anatômicas extremas. Por exemplo, pregas gástricas muito proeminentes com coloração avermelhada intensa e natural podem ser interpretadas equivocadamente pelo método como sinais de inflamação clínica, como na esofagite.
2. Apesar dos resultados expressivos obtidos na base Kvasir V1, a validação em bases externas ainda se faz necessária, pois o uso de dados provenientes de diferentes equipamentos e protocolos de aquisição é fundamental para comprovar a robustez e a aplicabilidade do método em ambientes clínicos reais.
3. Devido à alta complexidade computacional no treinamento das arquiteturas, não foi viável a execução de testes estatísticos mais profundos, como o teste *t* corrigido de Nadeau.

Os aspectos positivos discutidos nas etapas do método proposto contribuíram para que os resultados obtidos fossem comparáveis aos trabalhos encontrados na literatura. De modo geral, é possível perceber que métodos baseados em arquiteturas isoladas, ou que combinam apenas modelos da mesma natureza, apresentam limitações na captura simultânea de informações locais e globais. Nesse contexto, o método proposto, ao integrar técnicas de processamento de imagens com um *Ensemble Stacking* híbrido, foi capaz de superar essas limitações e alcançar resultados consistentes. Portanto, este estudo apresenta contribuições técnicas relevantes e se mostra como uma ferramenta promissora de apoio à decisão, com potencial para auxiliar profissionais de saúde e contribuir para o diagnóstico precoce de anomalias.

6 CONCLUSÃO

A análise automática de imagens endoscópicas tem sido amplamente investigada como forma de apoiar especialistas na detecção de anormalidades gastrointestinais. Nesse contexto, este trabalho apresentou um método computacional para a classificação binária de imagens endoscópicas em classes normal e anormal, com o objetivo de apoiar o processo de triagem de exames endoscópicos. O método proposto integra etapas de extração da ROI, pré-processamento com redução de SH e DA, além da classificação baseada em *Ensemble Stacking*, que combina arquiteturas de CNNs e ViT. Os resultados experimentais, incluindo análise de ablação e estatísticos, e estudos de caso, demonstraram a contribuição dessas etapas para o desempenho do método, que alcançou 98,12% de acurácia, 98,15% de precisão, 98,12% de sensibilidade, 98,23% de especificidade e F1-score de 98,13%, superando abordagens relatadas na literatura.

Como trabalhos futuros, planeja-se validar o método em bases de imagens externas, aplicar estratégias de balanceamento de classes, utilizar abordagens bayesianas para comparar os modelos, executar o teste de hipótese em todas as etapas do método e explorar a otimização automatizada de hiperparâmetros, visando elevar a capacidade de generalização dos modelos para futuras aplicações em ambientes clínicos reais.

6.1 Produções Científicas

A Tabela 8 apresenta os artigos diretamente relacionados ao método proposto. Além disso, a Tabela 9 lista os trabalhos publicados em outras aplicações durante o período da graduação.

Tabela 8 – Produções científicas em relação ao método proposto.

Artigo	**Tipo**	**Qualis**	**Status**
Anomalies diagnostic in endoscopic images using Deep Learning Ensemble models. Em: Brazilian Conference on Intelligent Systems (BRACIS). Ano: 2024.	Congresso	A4	Publicado
Diagnósticos de Anomalias em Imagens Endoscópicas usando Deep Learning. Em: 1ª Semana de Ensino, Pesquisa, Extensão e Cultura (SEPEC). Ano: 2024.	Evento local	-	Publicado
Ensemble Stacking de CNNs e *Vision Transformers* para Classificação de Anormalidades em Imagens Endoscópicas Em: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). Ano: 2026.	Congresso	A4	Submetido

Fonte: Elaborado pelo autor.

Tabela 9 – Outras produções científicas durante a graduação.

Trabalho	**Tipo**	**Qualis**	**Status**
Painel de Monitoramento Contínuo e Remoto de Nível D'água de Cisterna para Gestão de Solução Alternativa de Abastecimento D'água. Em: XXVI Simpósio Brasileiro de Recursos Hídricos (SBRH). Ano: 2025.	Congresso	-	Publicado
PyGo. Ano: 2024.	Registro de software	-	Publicado

Fonte: Elaborado pelo autor.

REFERÊNCIAS

ACKERMAN, L. V.; GOSE, E. E. *et al.* Breast lesion classification by computer and xeroradiograph. **Cancer**, v. 30, n. 4, p. 1025–1035, 1972. Um dos primeiros relatos de classificação computadorizada de lesões mamográficas. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/4562502/>.

AHLAWAT, R.; HOILAT, G. J.; ROSS, A. B. **Esophagogastroduodenoscopy**. Treasure Island (FL): StatPearls Publishing, 2023. StatPearls [Internet]. Last update: Aug 8, 2023. Bookshelf ID: NBK532268.

AHMAD, I.; SHANG, F.; PATHAN, M. S.; WAJAHAT, A.; KIM, Y.-S. Dual-stream hybrid architecture with adaptive multi-scale boundary-aware mechanisms for robust urban change detection in smart cities. **Scientific Reports**, v. 15, 08 2025.

AJLAN, R. S.; DESAI, A. A.; MAINSTER, M. A. Endoscopic vitreoretinal surgery: principles, applications and new directions. **International Journal of Retina and Vitreous**, BioMed Central, v. 5, n. 1, p. 15, 2019.

ALAGAPPAN, M. *et al.* Artificial intelligence in gastrointestinal endoscopy: The future is almost here. **National Library of Medicine**, 2018.

AVANZO, M.; STANCANELLO, J.; PIRRONE, G.; DRIGO, A.; RETICO, A. The evolution of artificial intelligence in medical imaging. **International Journal / review (PMC)**, 2024. Artigo de revisão que reconstrói a linha do tempo desde as primeiras aplicações computacionais até o uso atual de deep learning. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11545079/>.

AWE, O. O.; OPATEYE, G. O.; JOHNSON, C. A. G.; TAYO, O. T.; DIAS, R. Weighted hard and soft voting ensemble machine learning classifiers: Application to anaemia diagnosis. In: **Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana, 2022**. [S.l.]: Springer, 2024. p. 351–374.

AYAN, E. Classification of gastrointestinal diseases in endoscopic images: Comparative analysis of convolutional neural networks and vision transformers. **Journal of the Institute of Science and Technology**, v. 14, n. 3, p. 988–999, 2024.

AZZOUZ, L. L.; SHARMA, S. **Physiology, Large Intestine**. [S.l.]: StatPearls Publishing, 2023. StatPearls [Internet]. Bookshelf ID: NBK507857. Last update: Jul 31, 2023.

CHAUDHRY, S. R.; BORDONI, B. **Anatomy, Thorax, Esophagus**. 2023. StatPearls [Internet]. NCBI Bookshelf. Last update: July 24, 2023.

CHIRAS, D. D. **Human body systems: Structure, function, and environment**. [S.l.]: Jones & Bartlett Publishers, 2013.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

CRUZ, L. B. d. *et al.* Interferometer eye image classification for dry eye categorization using phylogenetic diversity indexes for texture analysis. **Computer Methods and Programs in Biomedicine**, v. 188, p. 105269, 2020.

CRUZ, L. B. d. *et al.* Kidney tumor segmentation from computed tomography images using deeplabv3+ 2.5 d model. **Expert Systems with Applications**, v. 192, p. 116270, 2022.

DELLON, E. S.; MUIR, A. B.; KATZKA, D. A.; SHAH, S. C.; SAUER, B. G.; ACEVES, S. S.; FURUTA, G. T.; GONSALVES, N.; HIRANO, I. Acg clinical guideline: Diagnosis and management of eosinophilic esophagitis. **The American Journal of Gastroenterology**, v. 120, p. 31–59, 2025.

DEMIRBA *et al.* Spatial-attention convmixer architecture for classification and detection of gastrointestinal diseases using the kvasir dataset. **Health Information Science and Systems**, Springer, v. 12, n. 1, p. 32, 2024.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **The Journal of Machine Learning Research**, JMLR. org, v. 7, p. 1–30, 2006.

DINIZ, J. O. B. *et al.* Heart segmentation in planning ct using 2.5 d u-net++ with attention gate. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, v. 11, n. 3, p. 317–325, 2023.

DOI, K. Computer-aided diagnosis in medical imaging: Historical review and current status. **Physics in Medicine and Biology (historical review collections) / Radiographics summaries**, 2009. Revisão histórica sobre a evolução do CAD em radiologia. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1955762/>.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An image is worth 16x16 words: Transformers for image recognition at scale. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2021.

DUDA, R. **Pattern classification and scene analysis**. New York, London, Sydney, Toronto: A Wiley-Interscience Publication, 1973.

FEUERSTEIN, J. D.; RAKOWSKY, S.; SATTLER, L.; YADAV, A.; FOROMERA, J.; GROSSBERG, L.; CHEIFETZ, A. S. Meta-analysis of dye-based chromoendoscopy studies. **Gastrointestinal Endoscopy**, 2019. Disponível em: <https://www.giejournal.org/article/S0016-5107(19)31602-5/fulltext>.

GE, H.; ZHOU, X.; WANG, Y.; XU, J.; MO, F.; CHAO, C.; ZHU, J.; YU, W. Development and validation of deep learning models for the multiclassification of reflux esophagitis based on the los angeles classification. **Journal of Healthcare Engineering**, v. 2023, p. 7023731, 2023. PMCID: PMC9966565.

GONZALEZ, R. C.; WOODS, R. E. **Processamento Digital de Imagens**. 3. ed. São Paulo: Pearson Prentice Hall, 2010.

GUO, Y.; LIU, Y.; OERLEMANS, A.; LAO, S.; WU, S.; LEW, M. S. Deep learning for visual understanding: A review. **Neurocomputing**, v. 187, p. 27–48, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. [S.l.]: Springer Science Business Media, 2009.

HE, K. *et al.* Deep residual learning for image recognition. **arXiv**, 2015.

Hospicenter. **Sistema de Vídeo Endoscopia SVE-100 — Argus (página de produto)**. n.d. Página da loja / ficha técnica do produto. Imagem: captura de tela do site feita pelo autor em 05 set. 2025. Disponível em: <https://www.hospicenter.com.br/sinais-vitais/monitores/sistema-de-video-endoscopia-sve-100-argus?srsltid=AfmBOoocmx4CypbnPhfzTsBvTW9ZI4Uu3j4wPWJoBijLKjf3osZ3gTBV>.

HUA, K. L.; HSU, C. H.; HIDAYATI, S. C.; CHENG, W. H.; CHEN, Y. J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. **OncoTargets and Therapy**, v. 8, p. 2015–2022, 2015.

HUBEL, D. H.; WIESEL, T. N. Early exploration of the visual cortex. **Neuron**, v. 20, n. 3, p. 401–412, 1998.

HUSSAIN, T.; SHOUNO, H.; HUSSAIN, A.; HUSSAIN, D.; ISMAIL, M.; MIR, T. H.; HSU, F. R.; ALAM, T.; AKHY, S. A. Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification. **IEEE Access**, IEEE, v. 13, p. 54040–54068, 2025.

ILIC, M.; ILIC, I. Epidemiology of stomach cancer. **World Journal of Gastroenterology**, v. 28, n. 12, p. 1187, 2022.

INCA, I. N. de C. **Estimativa 2023: incidência de câncer no Brasil**. Rio de Janeiro, RJ, 2022.

ISLAM NEAL C. PATEL, D. L.-H. R. S.; NGUYEN, C. C. Gastric polyps: A review of clinical, endoscopic, and pathologic features. **Gastroenterology Research and Practice / Public access (PMC)**, 2014. PMCID: PMC3992058. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3992058/>.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398.

JÚNIOR, D. A. D. *et al.* Automatic method for classifying covid-19 patients based on chest x-ray images, using deep features and pso-optimized xgboost. **Expert Systems with Applications**, v. 183, p. 115452, 2021.

KANG, S. H.; HYUN, J. J. Preparation and patient evaluation for safe gastrointestinal endoscopy. **Clinical Endoscopy**, 2013. Accessed: 2025-09-04. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678055/>.

KITTLER, J.; HATEF, M.; DUIN, R. P.; MATAS, J. On combining classifiers. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 20, n. 3, p. 226–239, 1998.

KOH, P. *et al.* Diagnostic utility of upper and lower gastrointestinal endoscopy for the diagnosis of acute graft-versus-host disease in children following stem cell transplantation: A 12-year experience. **Pediatric Transplantation**, v. 25, n. 7, p. e14046, 2021.

KOHAVI, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI)**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, p. 436–444, 2015.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LODWICK, G. S.; HAUN, C. L.; SMITH, W. E.; KELLER, R. F.; ROBERTSON, E. D. Computer diagnosis of primary bone tumors. **Radiology**, v. 80, p. 273–275, 1963. Precursor histórico do uso de computadores na interpretação de radiografias. Disponível em: <https://pubs.rsna.org/doi/pdf/10.1148/80.2.273>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.]: University of California Press, 1965. v. 1, p. 281–297.

MCCARTHY, J.; MINSKY, M.; ROCHESTER, N.; SHANNON, C. E. **A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence**. 1956. Dartmouth College (research proposal).

NEWELL, A.; SHAW, J. C.; SIMON, H. A. **The Logic Theorist: A Program That Simulates the Problem-Solving Behavior of a Human**. 1956. Technical report / presented work (1956).

NOFFSINGER, A. E. Serrated polyps and colorectal cancer: New pathway to malignancy. **Annual Review of Pathology: Mechanisms of Disease**, 2008.

PAK, A.; ZIYADEN, A.; TUKESHEV, K.; JAXYLYKOVA, A.; ABDULLINA, D. Comparative analysis of deep learning methods of detection of diabetic retinopathy. **Cogent Engineering**, v. 7, 08 2020.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. **arXiv preprint arXiv:1712.04621**, 2017.

POGORELOV, K. *et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: **ACM**. [S.l.: s.n.], 2017.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81–106, 1986.

RAO, J.-N.; WANG, J. Y. Intestinal architecture and development. In: **Regulation of Gastrointestinal Mucosal Growth**. San Rafael, CA: Morgan & Claypool Life Sciences, 2010. Bookshelf ID: NBK54098.

RISKA, S.; SULISTYO, D.; MAHARANI, F. High-accuracy classification of banana varieties using resnet-50 and densenet-121 architectures. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 39, p. 322, 07 2025.

ROGLER, G. Chronic ulcerative colitis and colorectal cancer. **Cancer Letters**, 2014.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958.

RUBIN, D. T.; ANANTHAKRISHNAN, A. N.; SIEGEL, C. A.; SAUER, B. G.; LONG, M. D. Acg clinical guideline: Ulcerative colitis in adults. **The American Journal of Gastroenterology**, v. 114, n. 3, p. 384–413, 2019. PDF copy downloaded from AJG (07/2019).

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533–536, 1986.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959.

SEHMUS, A. Ensemble-based deep transfer learning for robust gastrointestinal endoscopy image classification. **Balkan Journal of Electrical and Computer Engineering**, v. 13, n. 1, 2025.

SHANNON, C. E. Programming a computer for playing chess. **Philosophical Magazine**, v. 41, p. 256–275, 1950.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–48, 2019.

SHRESTHA, A.; MAHMOOD, A. Review of deep learning algorithms and architectures. **IEEE Access**, v. 7, p. 53040–53065, 2019.

SIDDIQUI, S.; KHAN, J. A.; ALGAMDI, S. Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification. **PeerJ Computer Science**, v. 11, p. e2809, 2025.

SIMADIBRATA, D. M.; LESMANA, E.; FASS, R. Role of endoscopy in gastroesophageal reflux disease. **Clinical Endoscopy**, v. 56, n. 6, p. 681–692, 2023. PMCID: PMC10665616.

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

STONE, M. Cross-validatory choice and assessment of statistical predictions. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 111–133, 1974.

SUBEDI *et al.* Classification of endoscopy and video capsule images using cnn-transformer model. **arXiv preprint arXiv:2408.10733**, 2024.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International conference on machine learning (ICML)**. [S.l.], 2019. p. 6105–6114.

THAMBAWITA, V.; STRüMKE, I.; HICKS, S. A.; HALVORSEN, P.; PARASA, S.; RIEGLER, M. A. Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. **Diagnostics**, v. 11, n. 12, 2021. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/11/12/2183>.

TURING, A. M. Computing machinery and intelligence. **Mind**, v. 59, n. 236, p. 433–460, 1950.

UNGARO, R.; MEHANDRU, S.; ALLEN, P. B.; PEYRIN-BIROULET, L.; COLOMBEL, J.-F. Ulcerative colitis. **The Lancet**, v. 389, n. 10080, p. 1756–1770, 2017. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/27914657/>.

UNIFAL-MG – Histologia Interativa. **Sistema Digestório – Histologia Interativa**. n.d. Página web. Acessado em: 04 set. 2025. Disponível em: <https://www.unifal-mg.edu.br/histologiainterativa/sistema-digestorio/>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. v. 30.

VIANA, P. da S.; CRUZ, L. B. da; JR, D. A. D.; DINIZ, J. O. B. Anomalies diagnostic in endoscopic images using deep learning ensemble models. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2024. p. 110–124.

WANG, W.; XIE, E.; LI, X.; FAN, D.-P.; SONG, K.; LIANG, D.; LU, T.; LUO, P.; SHAO, L. Pvtv2: Improved baselines with pyramid vision transformer. **Computational Visual Media**, Springer, v. 8, n. 3, p. 415–424, 2022.

WOLPERT, D. H. Stacked generalization. **Neural Networks**, Elsevier, v. 5, n. 2, p. 241–259, 1992.

WOLPERT, D. H. Stacked generalization. **Neural networks**, Elsevier, v. 5, n. 2, p. 241–259, 1992.

ZHANG, C.; LIU, Y.; WANG, K.; TIAN, J. Specular highlight removal for endoscopic images using partial attention network. **Physics in Medicine & Biology**, IOP Publishing, v. 68, n. 22, p. 225006, 2023.

ZHOU, Z.-H. **Ensemble Methods: Foundations and Algorithms**. [S.l.]: Chapman and Hall/CRC, 2012.

ANEXO

DECLARAÇÃO

Declaro para os devidos fins que este Trabalho de Conclusão de Curso (Monografia/Tese/Dissertação), escrito sob minha orientação, está em versão final, de acordo com as solicitações realizadas pela banca examinadora.

Informo também que procedi à revisão final do texto, constatando que atende às especificações das normas da ABNT para apresentação de trabalhos acadêmicos da UFCA, no que diz respeito ao conteúdo e à formatação.

07/04/2026

gov.br
Documento assinado digitalmente
LUANA BATISTA DA CRUZ
Data: 07/04/2026 17:17:15-0300
Verifique em https://validar.iti.gov.br