Ensemble Stacking de CNNs e *Vision Transformers* para Classificação de Anormalidades em Imagens Endoscópicas

Pedro da S. Viana[1], **Luana B. da Cruz**[1], **João O. B. Diniz**[2], **Nelson C. Sandes**[1]

[1]Laboratório de Inteligência Computacional Aplicada (LICA)
Universidade Federal do Cariri (UFCA)

[2]Fábrica de Inovação
Instituto Federal do Maranhão (IFMA)

pedro.viana@aluno.ufca.edu.br, luana.batista@ufca.edu.br

joao.bandeira@ifma.edu.br, nelson.sandes@ufca.edu.br

Abstract. *The massive analysis of endoscopic examinations overwhelms specialists in identifying gastrointestinal abnormalities. To optimize this screening process, this work proposes a binary classification method (normal and abnormal) using the Kvasir V1 image database. 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, distinguishing itself as a robust tool for medical diagnostic support.*

Resumo. *A análise massiva de exames endoscópicos sobrecarrega especialistas na identificação de anormalidades gastrointestinais. Para otimizar essa triagem, este trabalho propõe um método de classificação binária (normal e anormal) utilizando a base de imagens Kvasir V1. 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%, destacando-se como uma robusta ferramenta de suporte ao diagnóstico médico.*

1. Introdução

O trato gastrointestinal é suscetível a diversas anormalidades, como pólipos e colite ulcerativa [Rogler 2014, Chiras 2013]. Essas alterações podem atuar como lesões precursoras e fatores de risco para neoplasias agressivas, como os cânceres gástrico e colorretal. No Brasil, no ano de 2023, o câncer de estômago ocupou o quinto lugar entre as neoplasias malignas, enquanto o câncer colorretal foi a terceira neoplasia mais comum no país [de Câncer INCA 2022]. Nesse contexto, a identificação precoce dessas anormalidades torna-se indispensável.

Para isso, procedimentos visuais como a endoscopia digestiva alta e baixa (colonoscopia) representam o padrão-ouro de avaliação [Ilic and Ilic 2022]. No entanto, a análise meticulosa do grande volume de dados gerados por esses exames impõe uma

carga significativa aos especialistas. Visando mitigar esse gargalo operacional e otimizar a triagem médica, sistemas de Detecção Assistida por Computador (CAD) e Diagnóstico Assistido por Computador (CADx) têm sido amplamente adotados, contribuindo para acelerar a análise e aprimorar a precisão do diagnóstico final [Gonçalves et al. 2024, Alvino et al. 2025].

Neste contexto, este trabalho propõe um método para classificação binária de imagens endoscópicas em classes normal e anormal, com foco no auxílio à triagem de exames gastrointestinais, com as seguintes contribuições: (1) combina Redes Neurais Convolucionais (CNNs) e *Vision Transformers* por meio de *Ensemble Stacking*, explorando a complementaridade entre representações locais e globais das imagens; (2) utiliza a extração da Região de Interesse (ROI) para remover regiões irrelevantes da imagem; e (3) aplica técnicas de pré-processamento baseadas na redução de *Specular Highlights* (SH) e aplicação de *Data Augmentation* (DA) para melhorar a qualidade das características extraídas pelos modelos. Essas estratégias contribuem para o desenvolvimento de uma abordagem voltada ao apoio à triagem automática de exames endoscópicos.

2. Trabalhos Relacionados

A literatura apresenta diversos trabalhos que utilizam aprendizado profundo para a classificação multiclasse de doenças gastrointestinais em imagens endoscópicas. A seguir, são apresentados alguns estudos relevantes.

No contexto da classificação de doenças gastrointestinais, [Ayan 2024] propuseram uma abordagem comparativa de aprendizado profundo 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%.

Já no trabalho [Demirbaş 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 per-

formance 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 de 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.

3. Materiais e Método Proposto

O método proposto compreende quatro etapas, validadas utilizando a base de imagens apresentada na subseção 3.1. 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 alta intensidade (SH) e a aplicação de técnicas de DA para diversificar a base. Na terceira etapa, diversas arquiteturas de CNNs e *Vision Transformers* foram treinadas e combinadas usando *Ensemble Stacking* para a construção do modelo preditivo. Por fim, na quarta etapa, foram calculadas as métricas de validação para mensurar a robustez do método. A Figura 1 ilustra o processo descrito.

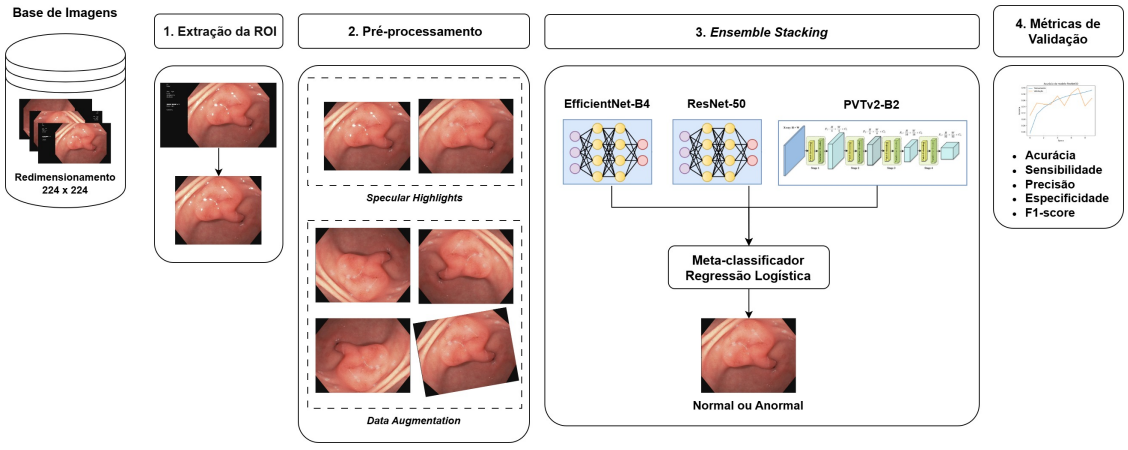

Figura 1. Visão geral do método proposto.

3.1. Base de Imagens

Utilizou-se a base pública Kvasir V1 [Pogorelov et al. 2017], composta por 4.000 imagens endoscópicas (com resoluções entre 730 x 576 e 1.920 x 1.072 *pixels*, no formato

JPG). Essa base foi escolhida devido ao seu amplo reconhecimento na literatura e ao fato de ter sido construída e anotada por especialistas médicos. Originalmente divididas em oito classes de 500 imagens cada, elas foram agrupadas para o escopo binário deste trabalho em: normais (Ceco Normal, Piloro Normal e Linha Z Normal; totalizando 1.500 imagens) e anormais (Pólipos Levantados Tingidos, Esofagite, Pólipos, Colite Ulcerativa e Margens de Ressecção Tingidas; totalizando 2.500 imagens). A Figura 2 apresenta exemplos das classes utilizadas neste estudo.

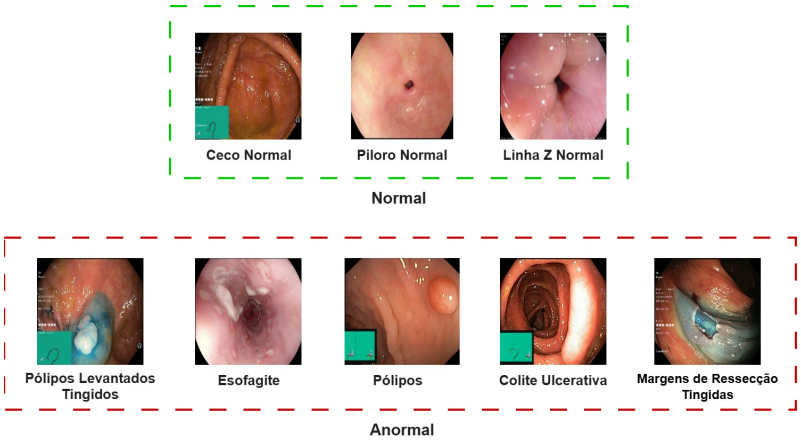

Figura 2. Separação das classes nas categorias normal e anormal.

Neste estudo, as imagens foram redimensionadas para 224 × 224 *pixels*, valor definido por representar um equilíbrio adequado entre custo computacional e desempenho. Para os experimentos neste estudo, a base foi particionada em conjuntos de treinamento (70%), validação (10%) e teste (20%) utilizando a técnica de amostragem estratificada.

3.2. 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 [da S. 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 3).

Figura 3. Etapas da extração da ROI.

3.3. 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 frequentemente gera SH que ocluem texturas relevantes e introduzem ruído visual, que podem degradar informações importantes do tecido e comprometer a extração de características discriminativas pelos modelos de aprendizado profundo. Para reduzir a interferência desses artefatos, realiza-se a segmentação dos *pixels* cuja luminância excede um limiar predefinido. Em seguida, aplica-se a técnica de *inpainting* através do algoritmo de Telea, configurado com um raio de vizinhança de 7 *pixels*, que reconstrói as áreas corrompidas interpolando informações da vizinhança adjacente, restaurando assim a região. No segundo, aplica-se DA com rotações aleatórias de até 10% e espelhamentos horizontal e/ou vertical (Figura 4). Esta estratégia diversifica a base de treinamento gerando novas variações das imagens originais, favorecendo a capacidade de generalização dos modelos.

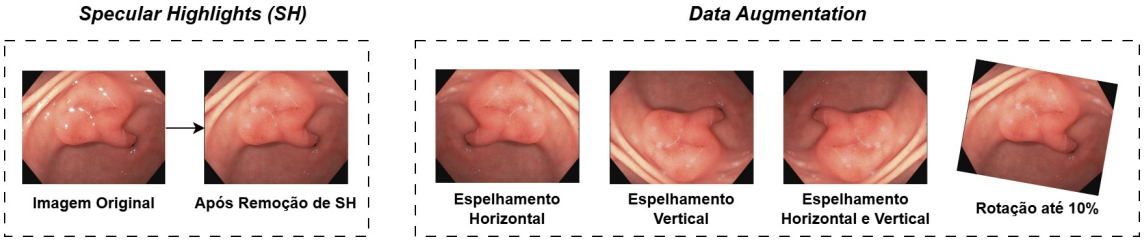

Figura 4. Processos da etapa de pré-processamento.

3.4. *Ensemble Stacking*

A etapa de classificação utiliza o *Ensemble Stacking* [Wolpert 1992] para extrair o máximo potencial da diversidade entre CNNs e *Vision Transformers*. A etapa de meta-classificação é executada por um algoritmo de Regressão Logística (RL) [Hosmer 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 *Pyramid Vision Transformer V2* (PVTv2-B2). Esse conjunto combina CNNs com diferentes capacidades de extração de características e uma arquitetura baseada em *Transformers*, 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 5).

3.5. Métricas de Validação

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

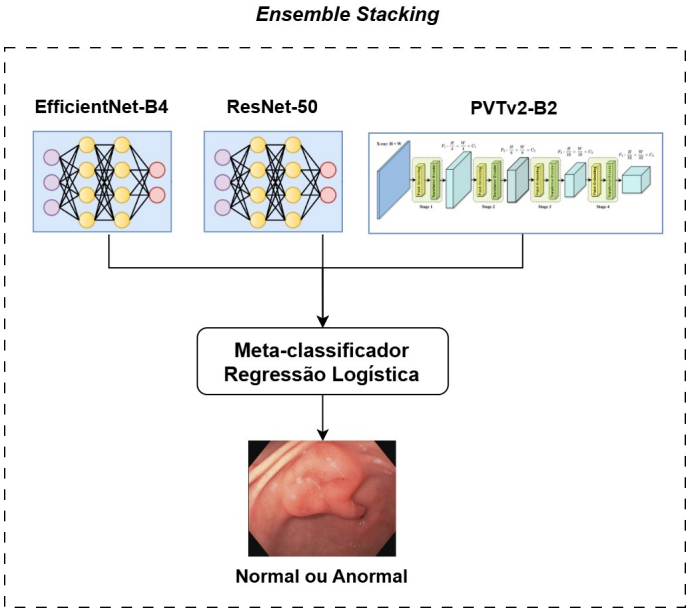

Figura 5. *Ensemble Stacking*.

[Duda 1973].

4. Resultados e Discussão

Nesta seção, 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.

4.1. 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*.

4.2. 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%**

Os resultados indicam de forma inequívoca melhorias em todas as métricas ao empregar a extração da ROI. Essa melhoria sugere 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.

4.3. Experimento com e sem Pré-processamento

Na Seção 3.3, 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%**

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 reduzindo o *overfitting*, resultando em classificações mais robustas pelo método proposto.

4.4. Experimento com Modelos Individuais

Nesta etapa, os modelos EfficientNet-B4, ResNet-50 e PVTv2-B2 foram avaliados individualmente utilizando as etapas de extração de ROI e pré-processamento. Os resultados estão sumarizados na Tabela 3.

Analisando a Tabela 3, observa-se que o PVTv2-B2 apresentou o melhor desempenho isolado entre as arquiteturas testadas. Esse resultado sugere que a capacidade dos

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

Experimento	ACC	PRE	SEN	ESP	F1-score
EfficientNet-B4	93,37%	93,66%	93,37%	93,42%	93,77%
ResNet-50	93,87%	94,15%	93,87%	94,30%	93,92%
PVTv2-B2	96,37%	96,48%	96,37%	96,63%	96,39%

Vision Transformers de focar na compreensão global do contexto visual é altamente eficaz para imagens endoscópicas. No entanto, as CNNs (EfficientNet-B4 e ResNet-50) mantiveram resultados competitivos, evidenciando sua proficiência na extração de características locais, o que motiva a integração das abordagens para compensar as fraquezas de cada modelo.

4.5. Comparação entre Métodos de *Ensemble*

Para integrar as predições dos modelos avaliados individualmente e explorar a complementaridade entre suas representações, foram investigadas 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%**

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 previsões 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 *Vision Transformers*, resultando em uma combinação mais eficaz de predições e melhorias consistentes nas métricas de desempenho.

4.6. 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,

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

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%**

composta pela redução de SH e pela aplicação de DA, resultou em novos avanços, refletindo uma melhoria na capacidade 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.

4.7. Estudos de Caso

Esta seção apresenta três estudos de caso qualitativos envolvendo a análise de classificações realizadas pelo método proposto, ilustrados na Figura 6. 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).

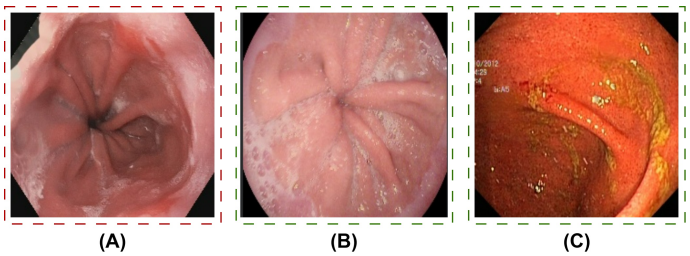

Figura 6. Estudos de caso. (A) Imagem normal classificada erroneamente como anormal; (B) Imagem anormal classificada corretamente como anormal; e (C) Imagem normal classificada corretamente como normal.

Ao analisar a Figura 6 (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 6 (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 6 (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.

4.8. Comparação com a Literatura

Esta seção propõe uma análise comparativa com os estudos discutidos na Seção 2. 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 6 resume os resultados alcançados na comparação.

Tabela 6. 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%
[Demirbaş 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%**

Diversos estudos na área empregam arquiteturas profundas para detectar anormalidades em exames endoscópicos. Trabalhos como [Ayan 2024] e [Demirbaş 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 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*. Diferente-

mente 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 *Vision Transformers*. 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 *Vision Transformers* 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. 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 *Vision Transformers*. Os resultados experimentais, incluindo análise de ablação 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 e explorar a otimização automatizada de hiperparâmetros, visando elevar a capacidade de generalização dos modelos para futuras aplicações em ambientes clínicos reais.

6. Agradecimentos

Este trabalho foi financiado pela Universidade Federal do Cariri (UFCA) e apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Ainda, os autores reconhecem o uso do LLM para verificação ortográfica, correção gramatical e assistência na tradução de termos específicos.

Referências

Alvino, A. et al. (2025). Abordagem baseada em deep features para diagnóstico de câncer seroso de ovário em imagens histopatológicas. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 401–412, Porto Alegre, RS, Brasil. SBC.

Awe, O. O. et al. (2024). 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*, pages 351–374. Springer.

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

Chiras, D. D. (2013). *Human body systems: Structure, function, and environment*. Jones & Bartlett Publishers.

da S. Viana, P. et al. (2024). Anomalies diagnostic in endoscopic images using deep learning ensemble models. In *Brazilian Conference on Intelligent Systems*, pages 110–124. Springer.

de Câncer INCA, I. N. (2022). Estimativa 2023: incidência de câncer no brasil. Technical report, INCA, Rio de Janeiro, RJ.

Demirbaş, A. A., Üzen, H., and Fırat, H. (2024). Spatial-attention convmixer architecture for classification and detection of gastrointestinal diseases using the kvasir dataset. *Health Information Science and Systems*, 12(1):32.

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

Gonçalves, J. et al. (2024). D.iagnóstica: Ferramenta cadx para diagnóstico de doenças pulmonares em imagens radiológicas. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 214–225, Porto Alegre, RS, Brasil. SBC.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hussain, T. et al. (2025). Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access*, 13:54040–54068.

Ilic, M. and Ilic, I. (2022). Epidemiology of stomach cancer. *World Journal of Gastroenterology*, 28(12):1187.

Pogorelov, K. et al. (2017). In *ACM*.

Rogler, G. (2014). Chronic ulcerative colitis and colorectal cancer. *Cancer Letters*.

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

Siddiqui, S., Khan, J. A., and Algamdi, S. (2025). Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification. *PeerJ Computer Science*, 11:e2809.

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

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.