Classificação Multi-Classe de Imagens Endoscópicas Gastrointestinais com GhostNetV3: Uma Abordagem Eficiente para Diagnóstico por Imagem

Beneilton Martins Leite[1], **Alexandre Cesar Pinto Pessoa**[1], **Carlos Eduardo Veras Gomes,** [1]**Darlan Bruno Pontes Quintanilha**[1]

[1]Núcleo de Computação Aplicada
Universidade Federal do Maranhão (UFMA)
Caixa Postal: 65.085-580 – São Luís, MA – Brasil

{beneilton.martins, alexandre.pessoa,
carlos.veras, dquintanilha}@nca.ufma.br

Abstract. *This work explores the use of the GhostNetV3 architecture for the multi-class classification of gastrointestinal endoscopic images, addressing the increasing need for clinical diagnostic support tools in environments with limited computational resources. The automatic analysis of digestive tract examinations is fundamental for early intervention, making the implementation of automated methodologies essential to assist in screening. This research utilizes the HyperKvasir dataset, focusing on a subset of 16 clinical classes, to train and evaluate the performance of this lightweight and efficient convolutional network fine-tuned from ImageNet. The results demonstrate that GhostNetV3 exhibits competitive performance compared to significantly larger Vision Transformers, highlighting the potential of compact CNNs in classifying medical pathologies without requiring heavy infrastructure. The application of the Random Erasing technique as a regularization strategy resulted in a Macro F1-score of 85.72% and an MCC of 91.29% using only 8.1 million parameters, showcasing the effectiveness and efficiency of the proposed approach.*

Resumo. *Este trabalho explora o uso da arquitetura GhostNetV3 para a classificação multiclasse de imagens endoscópicas gastrointestinais, abordando a crescente necessidade de ferramentas de suporte ao diagnóstico clínico em ambientes com recursos computacionais limitados. A análise automática de exames do trato digestivo é fundamental para intervenções precoces, tornando essencial a implementação de metodologias automatizadas para auxiliar na triagem. Esta pesquisa utiliza o dataset HyperKvasir, com foco em um subconjunto de 16 classes clínicas, para treinar e avaliar o desempenho dessa rede convolucional leve e eficiente, ajustada a partir de pesos pré-treinados no ImageNet. Os resultados demonstram que a GhostNetV3 apresenta desempenho competitivo quando comparada a Vision Transformers significativamente maiores, evidenciando o potencial de CNNs compactas na classificação de patologias médicas sem a necessidade de infraestrutura computacional robusta. A aplicação da técnica de Random Erasing como estratégia de regularização resultou em um F1-Score Macro de 85,72% e um MCC de 91,29% utilizando apenas 8,1 milhões de parâmetros, demonstrando a eficácia e a eficiência da abordagem proposta.*

1. Introdução

Exames de imagem endoscopicos são utilizados no diagnóstico de patologias do trato gastrointestinal, como gastrites, úlceras pépticas, doença do refluxo gastroesofágico e neoplasias. Diferentes modalidades de aquisição, incluindo endoscopia convencional e cápsula endoscópica, produzem imagens com características visuais distintas. A interpretação desses exames é desafiadora devido à alta variabilidade das lesões, à presença de artefatos e à dependência da experiência do especialista [Wang et al. 2025]. Nesse contexto, métodos baseados em aprendizado profundo (*deep learning*) têm demonstrado elevada capacidade na identificação de padrões patológicos, contribuindo para o aumento da precisão diagnóstica e a redução da carga de trabalho clínico [Pandian 2025, Shobayo and Saatchi 2025]. Apesar desses avanços, ainda há uma lacuna em soluções que conciliem desempenho e eficiência computacional, especialmente em cenários com infraestrutura limitada.

Avanços recentes têm sido impulsionados por métodos de aprendizado autosupervisionado (*Self-Supervised Learning* – SSL), como *Masked Autoencoders* (MAE) [He et al. 2022], e por arquiteturas baseadas em atenção, como *Vision Transformers* (ViT) [Espantaléon-Pérez et al. 2023], que apresentam desempenho competitivo em tarefas de classificação médica [Aburass et al. 2025]. Entretanto, tais abordagens frequentemente envolvem elevado custo computacional [Aburass et al. 2025, Tang et al. 2025], com centenas de milhões de parâmetros e alta demanda de operações [Zhuang et al. 2025], limitando sua aplicação em cenários com restrições de hardware ou requisitos de inferência em tempo real [Choi et al. 2025].

Como alternativa, arquiteturas convolucionais eficientes permanecem relevantes. A GhostNet [Han et al. 2020] reduz redundâncias em mapas de características ao gerar *ghost features* por meio de operações lineares de baixo custo. Sua versão mais recente, GhostNetV3 [Liu et al. 2024], introduz melhorias de treinamento e eficiência, mantendo um número reduzido de parâmetros.

Neste trabalho, investigamos o uso da GhostNetV3, pré-treinada no ImageNet [Deng et al. 2009], para classificação multiclasse de imagens endoscópicas no dataset HyperKvasir (16 classes). O modelo é avaliado sob validação cruzada 5-fold, incorporando *Random Erasing* [Zhong et al. 2020] como estratégia de regularização para aumentar a robustez. Os resultados mostram que é possível alcançar desempenho competitivo com baixo custo computacional, evidenciando a viabilidade de arquiteturas convolucionais compactas em aplicações clínicas com recursos limitados.

2. Trabalhos Relacionados

A transferência de aprendizado a partir de grandes bases de dados de imagens em larga escala, como o ImageNet, tornou-se prática comum em tarefas de visão computacional aplicada à medicina [Deng et al. 2009]. Entretanto, imagens médicas apresentam características próprias, como menor variabilidade semântica e padrões visuais específicos, o que tem motivado o desenvolvimento de estratégias de aprendizado mais adequadas a esse domínio [Azizi et al. 2021]. No contexto de imagens endoscópicas, o conjunto de dados HyperKvasir tornou-se uma referência importante para avaliação de métodos de classificação.

Thambawita et al. (2022) investigaram o impacto da resolução da imagem no desempenho de redes convolucionais profundas no HyperKvasir. Utilizando arquiteturas

como DenseNet e ResNet, os autores demonstraram que o aumento da resolução de entrada pode melhorar significativamente o desempenho dos modelos, destacando a importância da qualidade das imagens para tarefas de classificação endoscópica.

Posteriormente, Espantaléon-Pérez et al. (2023) analisaram arquiteturas baseadas em mecanismos de atenção, incluindo MobileViT, CoAtNet, CMT e DaViT. Os autores observaram que artefatos visuais presentes no conjunto de dados podem induzir os modelos a aprender padrões espúrios, caracterizando o fenômeno conhecido como *shortcut learning* [Geirhos et al. 2020]. Seus experimentos também indicam que modelos baseados em atenção podem alcançar desempenho competitivo mesmo com número relativamente reduzido de parâmetros.

Mais recentemente, estratégias de aprendizado auto-supervisionado (*Self-Supervised Learning - SSL*) têm demonstrado grande potencial para aproveitar grandes volumes de dados não rotulados, comuns em bases médicas. Guo et al. (2024) propuseram o C-Mixup, uma abordagem que combina *curriculum learning* e *Mixup* ao arcabouço contrastivo SimSiam, pré-treinando o modelo em 100 mil imagens não rotuladas do HyperKvasir e ajustando-o nas 10 mil imagens rotuladas.

Em linha similar, Bravo et al. (2024) exploraram o uso de *Vision Transformers* (VIT) treinados com *Masked Autoencoders* (MAE) em imagens endoscópicas. Seus experimentos mostraram que o pré-treinamento auto-supervisionado no próprio domínio médico, seguido de *fine-tuning* supervisionado, produz representações mais robustas do que a simples transferência de pesos pré-treinados na ImageNet do ImageNet.

Os trabalhos analisados priorizam arquiteturas profundas ou modelos baseados em Transformers. Diferentemente destes métodos, este trabalho investiga o uso de uma CNN eficiente na tarefa de classificação multi-classe de patologias do trato gastrointestinal, e avaliando na base de dados HyperKvasir, utilizando 16 das classes disponíveis. Além da performance na tarefa de classificação, também são analisados aspectos relacionados ao custo computacional do modelo, aspecto relevante para aplicações em cenários clínicos com restrições computacionais.

3. Fundamentação Teórica

Nesta seção, serão apresentadas as técnicas que fundamentam a metodologia utilizada no trabalho.

3.1. GhostNet

A GhostNet é uma arquitetura baseada na hipótese de que muitos mapas de características gerados por convoluções são redundantes e podem ser aproximados por transformações lineares de baixo custo [Han et al. 2020]. Em vez de produzir diretamente todos os mapas de saída por meio de convoluções densas, o módulo *Ghost* decompõe essa operação em duas etapas.

Primeiramente, um subconjunto reduzido de mapas, denominado mapas intrínsecos, é gerado por uma convolução padrão (tipicamente 1×1), responsável por capturar relações entre canais. Em seguida, os mapas restantes, chamados *ghost feature maps*, são obtidos a partir desses mapas intrínsecos por meio de operações lineares baratas, implementadas como convoluções *depthwise* (i.e., convoluções com o parâmetro *groups* igual ao número

de canais de entrada, de modo que cada filtro opera sobre um único canal), usualmente com kernels 3×3. Essas operações atuam de forma independente em cada canal, não realizando mistura entre canais.

Formalmente, dado um número desejado de mapas de saída n, apenas $m = \frac{n}{s}$ mapas intrínsecos são gerados via convolução completa, enquanto os $(s-1)m$ mapas restantes são obtidos por transformações lineares baratas. Esse fator s, denominado *ghost ratio*, controla diretamente o compromisso entre custo computacional e capacidade representacional.

Dessa forma, enquanto uma convolução padrão possui custo proporcional a $C_{in} \cdot C_{out} \cdot k^2$, o módulo Ghost reduz esse custo ao decompor a operação em uma convolução densa sobre um número reduzido de canais, seguida por convoluções *depthwise*, cujo custo é proporcional a $C \cdot k^2$. Essa fatoração explora a redundância estrutural dos mapas de características, resultando em uma redução significativa de FLOPs e parâmetros, sem degradação relevante de desempenho.

A arquitetura completa é construída a partir de blocos *Ghost Bottleneck*, que combinam expansão via módulos *Ghost*, convoluções *depthwise* para captura espacial (com kernels tipicamente 3×3 ou 5×5), e projeção linear para redução dimensional, podendo ainda incorporar mecanismos de atenção como *Squeeze-and-Excitation*. A Figura 1 ilustra o módulo *Ghost*, unidade fundamental da arquitetura.

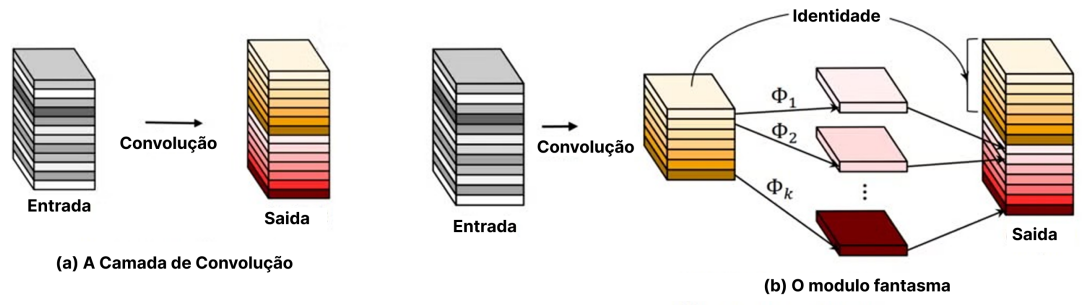

Figura 1. Estrutura do módulo *ghost* da GhostNet. Adaptado de [Han et al. 2020].

A GhostNetV3 [Liu et al. 2024] introduz melhorias principalmente no regime de treinamento de redes compactas, ao invés de alterações estruturais profundas. O modelo incorpora estratégias como reparametrização estrutural durante o treinamento e *knowledge distillation* adaptado para redes de baixa capacidade. Essas modificações resultam em um melhor equilíbrio entre eficiência computacional e desempenho preditivo, especialmente em cenários de *edge computing*.

3.2. *Random Erasing*

Random Erasing é uma técnica de aumento de dados que remove aleatoriamente regiões retangulares da imagem durante o treinamento, substituindo os pixels por valores aleatórios [Zhong et al. 2020]. Dada uma imagem I, uma região retangular R de área A é selecionada aleatoriamente, e seus pixels são substituídos por valores aleatórios o preenchimento é feito com valores uniformemente distribuídos no intervalo [0, 255] para

imagens de 8 bits, ou [0, 1] para tensores normalizados. A probabilidade de aplicação p e a proporção de área $A/(h \cdot w)$ são hiperparâmetros ajustáveis.

Essa técnica força o modelo a não depender exclusivamente de padrões locais específicos, incentivando o aprendizado de representações mais robustas e menos propensas a *overfitting*. É particularmente útil em domínios como imagens de exames endoscópicos, onde oclusões e artefatos são comuns. Nos experimentos conduzidos neste trabalho, foi aplicado o *Random Erasing* com probabilidade $0,5$, área removida entre 8% e 24% da imagem e proporção de aspecto entre $0,3$ e $3,3$, preenchendo a região removida com valores aleatórios.

3.3. Métricas de Avaliação

Devido ao desbalanceamento entre as classes, adotamos métricas menos sensíveis à predominância de classes majoritárias. Todas as métricas são calculadas no esquema *one-vs-rest* e reportadas como média macro quando aplicável.

F1-score: média harmônica entre precisão e sensibilidade (recall) por classe. A precisão (fração de predições corretas entre as amostras classificadas como positivas) e a sensibilidade (fração de positivos corretamente identificados) são combinados no F1 por classe, cuja média simples entre todas as classes resulta no F1 macro. Esta métrica equilibra os erros de falsos positivos e falsos negativos.

MCC (Matthews Correlation Coefficient): coeficiente que mede a qualidade da classificação para problemas multiclasses, calculado globalmente a partir da matriz de confusão. Varia de -1 (classificação inversa) a +1 (classificação perfeita) e é forte a desbalanceamentos por considerar todas as entradas da matriz de confusão.

Area sob a curva ROC (AUC): A AUC (*Area Under the Curve*) calculada no esquema *one-vs-rest*, com média macro. Para cada classe, a curva ROC (*Receiver Operating Characteristic*) é obtida comparando-a contra as demais, e a respectiva AUC reflete a capacidade do modelo em distinguir aquela classe das outras. A média macro das AUCs por classe fornece uma visão global do poder discriminativo do modelo, sendo resistente ao desbalanceamento por atribuir o mesmo peso a todas as classes.

Sensibilidade (recall) macro e especificidade macro: a sensibilidade mede a proporção de verdadeiros positivos corretamente identificados, enquanto a especificidade, definida como a proporção de verdadeiros negativos entre o total de negativos, mede a capacidade de identificar corretamente amostras de outras classes. Em conjunto, fornecem uma visão complementar do desempenho.

Acurácia: reportada apenas como referência global, dado seu viés em cenários desbalanceados.

4. Metodologia

A metodologia utilizada neste trabalho avalia a GhostNetV3 para classificação multi-classe de imagens endoscópicas do HyperKvasir. Nesta seção, são apresentados os detalhes relacionados à base de imagens utilizada, ao modelo empregado, ao processo de treinamento, às estratégias de aumento de dados e à análise de explicabilidade. A Figura 2 resume visualmente as etapas da metodologia adotada.

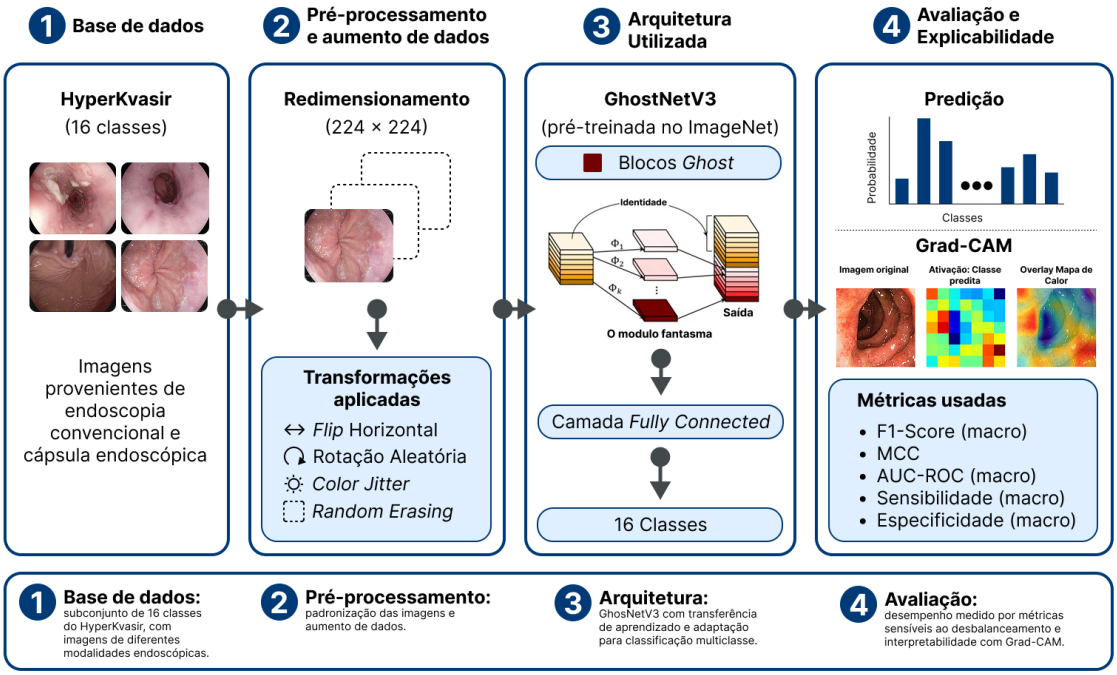

Figura 2. Fluxo metodológico do estudo: (1) seleção e divisão do conjunto Hyper-Kvasir em 5 *folds*, (2) aplicação de pré-processamento e aumento de dados, (3) treinamento da arquitetura GhostNetV3 para classificação das imagens e (4) análise de interpretabilidade utilizando Grad-CAM.

4.1. Dataset e Pré-processamento

A base de dados contém 10,662 imagens rotuladas em 23 classes [Borgli et al. 2020] . Para este trabalho, focamos no subconjunto de 16 classes, removendo classes com poucas amostras, seguindo a abordagem adotada por Guo et al (2024), excluindo aquelas com número de amostras inferior a 10% do tamanho da classe majoritária. Foi utilizada a separação disponibilizada no repositório da base de dados, que incluem divisões para validação cruzada de 2 e 5 *folds*. Para avaliar o modelo de forma mais confiável e reduzir o viés associado a uma única divisão treino-teste, empregamos validação cruzada com 5 *folds*. Os resultados são apresentados como média ± desvio padrão ao longo dos *folds*. A Figura 3 ilustra exemplos de imagens do dataset para diferentes classes.

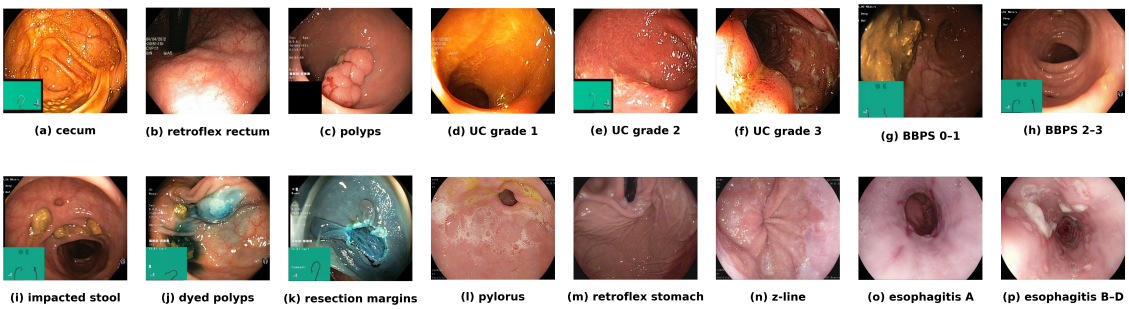

Figura 3. Recorte das 16 classes escolhidas entre as 23 totais do dataset

4.2. Arquitetura e Treinamento

Neste trabalho, foi utilizada a arquitetura GhostNetV3, com versão pré-treinada no ImageNet [Deng et al. 2009]. O modelo foi empregado na tarefa de classificação das 16 classes de patologias e achados anatômicos do trato gastrointestinal. A Tabela 1 apresenta o modelo utilizado, seu tamanho de entrada e o conjunto de dados no qual foi pré-treinado.

Tabela 1. Modelo CNN GhostNet utilizado neste estudo

Modelo	Tamanho de Entrada	Pré-treinado em	Parâmetros
GhostNetV3 [Liu et al. 2024]	224x224	ImageNet	8,1M

O pipeline de treinamento foi padronizado para todos os experimentos, alterando apenas a aplicação ou não do *Random Erasing*, garantindo assim que as configurações de treinamento e os hiperparâmetros fossem idênticos entre as diferentes configurações avaliadas. Além disso, essa abordagem assegura reprodutibilidade e comparabilidade entre os experimentos.

As imagens foram pré-processadas utilizando transformações comuns em redes neurais. Durante o treinamento, foi aplicado *random resized crop* para o tamanho de entrada do modelo (224×224), com escala no intervalo $[0,8,1,0]$ e razão de aspecto entre $[0,9,1,1]$. Em seguida, utilizou-se inversão horizontal aleatória com probabilidade de 0,5 e rotações aleatórias no intervalo de ±20°, ajustes de cor (*color jitter*), com variações de brilho e contraste de até 15%, saturação de até 10% e matiz de até 2%. Por fim, as imagens foram convertidas para tensores e normalizadas com base na média e no desvio padrão do ImageNet, de modo a manter a distribuição dos dados de entrada compatível com a utilizada no pré-treinamento dos modelos, o que contribui para maior estabilidade e melhor desempenho no *transfer learning*.

Para os experimentos com *Random Erasing*, foi adicionada a transformação *RandomErasing* com probabilidade de aplicação de 0,5, removendo regiões retangulares com área entre 8% e 24% da imagem e razão de aspecto entre 0,3 e 3,3, com preenchimento por valores aleatórios. Para a fase de validação, adotou-se um pré-processamento determinístico, consistindo em redimensionamento da imagem para 256 pixels no menor lado, seguido de *center crop* para 224×224 e normalização com os mesmos parâmetros utilizados no treinamento.

4.3. Treinamento dos Modelos

O modelo foi inicializado com os pesos pré-treinados no ImageNet. Foram estabelecidas 20 épocas por treinamento, e um tamanho de batch de 64. A técnica de treinamento utilizada foi a validação cruzada k-fold com 5 *folds*, seguindo os *splits* disponibilizados pelos autores da base de dados.

O otimizador utilizado foi AdamW [Loshchilov and Hutter 2019], com *learning rate* inicial de 6×10^{-4} e weight decay de 1×10^{-4}. Para otimizar o treinamento, empregamos a política de agendamento de taxa de aprendizado em um ciclo (1cycle policy), proposta por Smith and Topin (2017) , que permite o uso de taxas de aprendizado mais altas e convergência mais rápida, com fase inicial de aquecimento (*warmup*) correspondente

a aproximadamente 30% do treinamento. A função de perda foi a Cross Entropy, sem ponderação entre classes.

Para cada fold, o modelo foi treinado por 20 épocas, com seleção do melhor modelo baseada no maior F1 macro. Ao final do treinamento de cada fold, foram obtidas as predições para as imagens do respectivo conjunto de validação. Ao final dos cinco *folds*, essas predições foram agrupadas para calcular as métricas gerais de desempenho, obtendo-se a média e o desvio padrão entre os *folds*. Esse procedimento garantiu uma avaliação consistente e robusta do modelo testado.

4.4. Análise de Explicabilidade com Grad-CAM

Para interpretar as decisões do modelo e identificar as regiões mais relevantes para a classificação, foi implementada uma versão do Grad-CAM [Selvaraju et al. 2019] direcionada a última camada convolucional da GhostNetV3. O Grad-CAM utiliza os gradientes da classe alvo fluindo para esta camada para produzir um mapa de ativação que destaca as regiões importantes da imagem.

Para cada classe de interesse (pólipos, colite ulcerativa graus 1–3 e esofagite), selecionamos aleatoriamente amostras do conjunto de teste e geramos mapas de ativação sobrepostos à imagem original. A Figura 4 apresenta alguns exemplos de imagens do conjunto de dados.

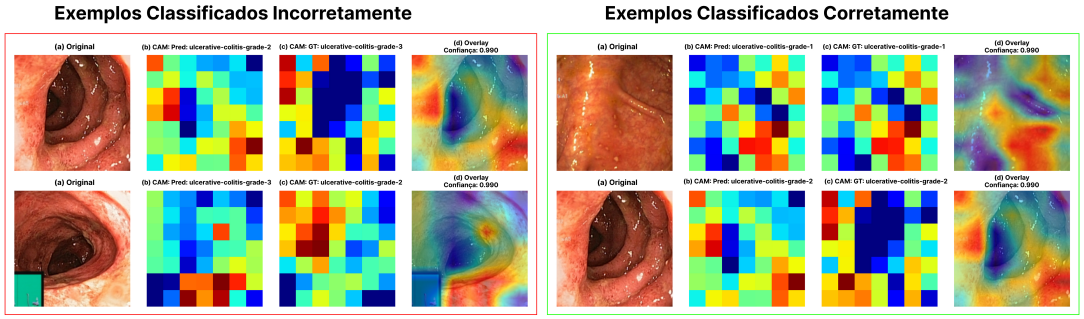

Figura 4. Exemplos de mapas de ativação Grad-CAM: (a) Imagem original, (b) Mapa de ativação para a classe predita, (c) Mapa de ativação para o *ground truth* (GT), (d) Sobreposição do mapa de ativação na imagem original

5. Resultados e Discussões

Nesta seção, apresentamos os resultados obtidos com o modelo GhostNetV3 nas configurações com e sem *Random Erasing*, destacando métricas relevantes para avaliar o desempenho do método na classificação das imagens endoscópicas.

5.1. Desempenho da GhostNetV3

A Tabela 2 mostra os resultados da GhostNetV3 com e sem *Random Erasing* (5-fold CV). O *Random Erasing* melhorou todas as métricas, com destaque para a redução do desvio padrão, indicando maior estabilidade entre os *folds*. O AUC-ROC próximo de 0,9952 e a especificidade alta mostram baixa taxa de falsos positivos.

Tabela 2. Comparação do desempenho do modelo GhostNetV3 com e sem o uso de *Random Erasing*. Os valores representam média ± desvio padrão obtidos na validação cruzada *5-fold*.

Métrica	Sem *Random Erasing*	Com *Random Erasing*
Acurácia	$0,9185 \pm 0,0058$	$\mathbf{0,9198 \pm 0,0055}$
F1-score (macro)	$0,8560 \pm 0,0133$	$\mathbf{0,8572 \pm 0,0116}$
AUC-ROC (macro)	$0,9943 \pm 0,0006$	$\mathbf{0,9952 \pm 0,0007}$
MCC	$0,9114 \pm 0,0063$	$\mathbf{0,9129 \pm 0,0060}$
Sensibilidade (macro)	$0,8550 \pm 0,0178$	$\mathbf{0,8563 \pm 0,0148}$
Especificidade (macro)	$0,9946 \pm 0,0004$	$\mathbf{0,9947 \pm 0,0004}$

5.2. Comparação com a Literatura

A Tabela 3 compara nosso melhor modelo (com *Random Erasing*) com resultados publicados. Vale observar que utilizamos validação cruzada de 5 *folds*, enquanto parte dos trabalhos comparados adotou 2 *folds* e alguns uma divisão própria de treino e validação. Diferenças no protocolo podem influenciar as métricas obtidas. Além disso, a comparação entre 16 e 23 classes não é direta, pois a redução do número de classes tende resultar em um aumento dos valores obtidos pelas métricas utilizadas.

Tabela 3. Comparação com trabalhos relacionados no HyperKvasir.

Modelo	Classes	Parâmetros	Acurácia	F1-Macro	MCC
DenseNet-161 [Thambawita et al. 2021]	23	$28,7M$	–	$63,5$	$90,0$
MobileViT Large [Espantaléon-Pérez et al. 2023]	23	$18M$	–	$63,4$	–
Ensemble (MobileViT) [Espantaléon-Pérez et al. 2023]	23	$20M$	–	$63,6$	–
ResNet50 SimSiam + C-Mixup [Guo et al. 2024]	16	$40,3M$	$88,92$	$73,39$	–
ViT-Large SSL (EN) [Bravo et al. 2025]	23	$304M$	$91,00$	$65,73$	$90,25$
ViT-Large SSL (EN) [Bravo et al. 2025]	16	$304M$	$93,74$	$89,23$	$93,19$
GhostNetV3 **(Trabalho proposto)**	**16**	**8,1M**	**91,98**	**85,72**	**91,29**

GhostNetV3 atinge Macro F1 de 85,72% em 16 classes, com 30–40× menos parâmetros que o ViT-Large SSL (304M) e F1-Macro 89,2%. Em relação ao ensemble de MobileViT (20M, 23 classes),o modelo utilizado tem menos da metade dos parâmetros e opera em um cenário de 16 classes, que é inerentemente mais simples, o que justifica o F1 mais alto.

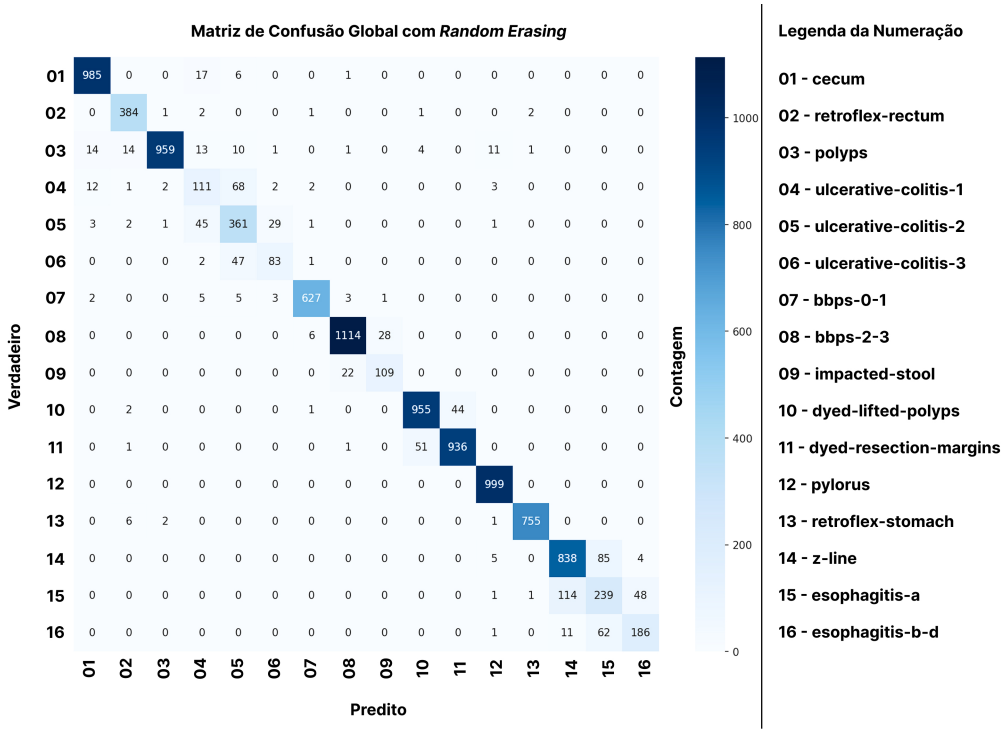

Figura 5. Matriz de confusão da GhostNetV3 para as 16 classes do HyperKvasir com *Random Erasing*

Uma análise da matriz de confusão, em conjunto com as visualizações dos mapas de ativação Grad-CAM, revela um padrão distinto: embora o modelo consistentemente focalize as regiões mucosas relevantes, ele apresenta dificuldade sistemática em discriminar os diferentes graus de colite ulcerativa. Esta limitação, portanto, não reside na localização das áreas de interesse, mas sim na sutil diferença visual entre os estágios da patologia, um desafio já documentado na literatura [Guo et al. 2024], [Bravo et al. 2025] e [Thambawita et al. 2021].

5.3. Eficiência Computacional

A GhostNetV3 tem cerca de 8.1M parâmetros e 0,83 GFLOPs. Comparado aos 304M e 59,65 Gflops do ViT-Large, 28,7M e 4,1 GFLOPs da ResNet50 ou as 28,7M e 7,73 Gflops da DenseNet161 a economia é substancial. Mesmo em relação ao MobileViT Large (18M), o presente modelo é mais leve. Isso torna a GhostNetV3 atraente para aplicações em tempo real ou em dispositivos com recursos limitados.

6. Conclusão

Este trabalho mostrou que a GhostNetV3, uma CNN leve pré-treinada no ImageNet e ajustada no HyperKvasir (16 classes), atinge Macro F1 de 85,72% e MCC de 91,29, com apenas 8,1M parâmetros. O desempenho fica próximo ao de modelos baseados em Transformers muito maiores (ViT-Large com 304M), evidenciando que arquiteturas convolucionais eficientes ainda são competitivas.

A comparação com trabalhos na literatura reforça a importância de protocolos consistentes (splits predefinidos, validação cruzada) para permitir comparações justas. Este

trabalho mostra que soluções leves são viáveis e devem ser consideradas em aplicações clínicas com restrição de recursos.

Ao avaliar o método proposto, identificaram-se algumas limitações que podem ser endereçadas em trabalhos futuros. Primeiramente, a abordagem atual não foi validada em outros conjuntos de dados de endoscopia, como Kvasir V2 e GastroVision. Além disso, o uso de pré-treinamento auto-supervisionado específico para o domínio de endoscopia poderia potencialmente melhorar o desempenho da GhostNetV3. Por fim, considerando aplicações práticas em dispositivos móveis, seria relevante explorar técnicas de compressão de modelo para reduzir o custo computacional sem comprometer a acurácia.

7. Agradecimentos

Os autores expressam sua gratidão às instituições brasileiras que contribuíram para o desenvolvimento desta pesquisa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA).

Referências

Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M., and Al-Kadi, O. (2025). Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *Journal of Imaging Informatics in Medicine*, 38(6):3928–3971.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, Montreal, Canada.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., and de Lange, T. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283.

Bravo, D., Ruano, J., Gómez, M., González, F. A., and Romero, E. (2025). Self-supervised learning for multi-category endoscopy classification and data quality evaluation using masked autoencoders. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Choi, M., Kim, S., and Lee, J. (2025). Edgesrie: A hybrid deep learning framework for real-time speckle reduction and image enhancement on portable ultrasound systems. *arXiv preprint arXiv:2507.03937*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA.

Espantaléon-Pérez, R. et al. (2023). Attention-based models for gastrointestinal endoscopy image classification. In *Computer Analysis of Images and Patterns (CAIP)*, volume 14185 of *Lecture Notes in Computer Science*, Cham. Springer.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Guo, H., Somayajula, S. A., Hosseini, R., and Xie, P. (2024). Improving image classification of gastrointestinal endoscopy using curriculum self-supervised learning. *Scientific Reports*, 14(1):6100.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1580–1589, Seattle, WA, USA.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, New Orleans, LA, USA.

Liu, Z. et al. (2024). Ghostnetv3: Exploring training strategies for compact models. *arXiv preprint arXiv:2404.xxxxx*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization.

Pandian, V. (2025). A comprehensive survey of deep learning methods in gastro-intestinal wireless capsule endoscopy images. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70052.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Shobayo, O. and Saatchi, R. (2025). Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. *Diagnostics*, 15(9).

Smith, L. N. and Topin, N. (2017). Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*.

Tang, F., Yao, Q., Ma, W., Wu, C., Jiang, Z., and Zhou, S. K. (2025). Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation.

Thambawita, V., Strümke, I., Hicks, S. A., Halvorsen, P., Parasa, S., and Riegler, M. A. (2021). Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. *Diagnostics*, 11(12):2183.

Wang, Y.-Y., Liu, B., and Wang, J.-H. (2025). Application of deep learning-based convolutional neural networks in gastrointestinal disease endoscopic examination. *World Journal of Gastroenterology*, 31(36):111137.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, New York, NY, USA.

Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., and Chen, H. (2025). Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE Transactions on Medical Imaging*, 44(11):4226–4238.