UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Daniel Lopes Soares Lima

Classificação de Imagens de Exames de Endoscopia por Cápsula Utilizando Transformers

São Luís
2023

Daniel Lopes Soares Lima

Classificação de Imagens de Exames de Endoscopia por Cápsula Utilizando Transformers

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Anselmo Cardoso de Paiva

Coorientador: Prof. Dr. António Manuel Trigueiros da Silva Cunha

São Luís - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Lima, Daniel Lopes Soares.
Classificação de Imagens de Exames de Endoscopia por Cápsula Utilizando Transformers / Daniel Lopes Soares Lima. - 2023.
57 f.

Coorientador(a): António Manuel Trigueiros da Silva Cunha.
Orientador(a): Anselmo Cardoso de Paiva.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luís, MA, 2023.

1. Classificação. 2. Endoscopia por cápsula. 3. Transformers. 4. Trato Gastrointestinal. I. Cunha, António Manuel Trigueiros da Silva. II. Paiva, Anselmo Cardoso de. III. Título.

Daniel Lopes Soares Lima

Classificação de Imagens de Exames de Endoscopia por Cápsula Utilizando Transformers

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho aprovado. São Luís - MA, 24 de Março de 2023:

Prof. Dr. Anselmo Cardoso de Paiva
Orientador
Universidade Federal do Maranhão

Prof. Dr. António Manuel Trigueiros da Silva Cunha
Coorientador
Universidade de Trás-os-Montes e Alto Douro

Prof. Dr. Darlan Bruno Pontes Quintanilha
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Augusto Marques Ferreira da Silva
Examinador Externo
Universidade de Aveiro

São Luís - MA
2023

À minha família, pelo apoio, suporte e constante incentivo ao estudo.

Agradecimentos

A Deus, pela força e perseverança para a conquista dos meus objetivos.

À minha esposa Lélia, e aos meus filhos Lucas e Laura, pelo encorajamento, confiança e amor incondicionais.

Aos meus professores do Programa de Pós-Graduação em Ciência da Computação, e em especial ao meu orientador Anselmo Paiva e ao meu coorientador António Cunha, pela oportunidade, conhecimentos transmitidos, disponibilidade e motivação para a conclusão desta etapa.

Aos meus colegas do Núcleo de Computação Aplicada, principalmente aos amigos Alan Lima e Alexandre Pessoa, sempre disponíveis e solícitos na sugestão de soluções para os problemas que se apresentaram durante a pesquisa.

Aos meus companheiros de trabalho da Diretoria de Gestão de Tecnologia da Informação da Reitoria do Instituto Federal de Educação, Ciência e Tecnologia do Maranhão, pelo apoio e incentivo.

E por fim a todos, que direta ou indiretamente, contribuíram para a realização deste trabalho.

"A persistência é o caminho do êxito."

(Charles Chaplin)

Resumo

As doenças inflamatórias intestinais apresentam alta taxa de incidência na população, sendo umas das principais causas de internação hospitalar. Os vídeos obtidos por meio de cápsulas endoscópicas são essenciais para o diagnóstico de anomalias no trato gastrointestinal. Porém, devido à sua duração, que pode chegar a 10 horas, demandam grande atenção do especialista médico em sua análise. Técnicas de aprendizado de máquina têm sido aplicadas com sucesso no desenvolvimento de sistemas de diagnóstico auxiliados por computador desde a década de 1990. Na última década as Redes Neurais Convolucionais (CNNs) tornaram-se modelo de grande sucesso para reconhecimento de padrões em imagens. As CNNs usam convoluções para extrair características dos dados analisados, operando em uma janela de tamanho fixo e, portanto, tendo problemas para capturar relacionamentos em nível de pixel considerando os domínios espacial e temporal. *Transformers*, por sua vez, usam mecanismos de atenção, onde os dados são estruturados em um espaço vetorial que pode agregar informações de dados adjacentes para determinar o significado em um determinado contexto. Este trabalho propõe um método computacional para análise de imagens extraídas de vídeos obtidos por cápsulas endoscópicas, usando uma arquitetura baseada em *Transformers*, visando auxiliar o especialista médico no diagnóstico de anormalidades do trato gastrointestinal. A metodologia proposta foi aplicada em 41511 imagens WCE do *dataset* Kvasir-Capsule. Nos experimentos realizados para a classificação de 11 classes, os melhores resultados foram alcançados pelo modelo DeiT, que registrou taxas médias de 99,75% de acurácia, 98,17% de precisão, 98,31% de sensibilidade e 98,06% de *f1-score*.

Palavras-chave: Trato Gastrointestinal, WCE, Classificação, Transformers, ViT, DeiT.

Abstract

Inflammatory bowel diseases have a high incidence rate in the population, being one of the leading causes of hospitalization. Videos obtained through endoscopic capsules are essential for evaluating anomalies in the gastrointestinal tract. However, due to their duration, which can reach 10 hours, they demand great attention from the medical specialist in their analysis. Machine learning techniques have been successfully applied in developing computer-aided diagnostic systems since the 1990s, where Convolutional Neural Networks (CNNs) have become very successful for pattern recognition in images. CNNs use convolutions to extract features from the analyzed data, operating in a fixed-size window and thus having problems capturing pixel-level relationships considering the spatial and temporal domains. Otherwise, Transformers use attention mechanisms, where data is structured in a vector space that can aggregate information from adjacent data to determine meaning in a given context. This work proposes a computational method for analyzing images extracted from videos obtained by endoscopic capsules, using a transformer-based model that helps diagnose of gastrointestinal tract abnormalities. The proposed methodology was applied on 41511 WCE images from the Kvasir-Capsule dataset. In the experiments performed for the classification task of 11 classes, the best results were achieved by the DeiT model, which registered average rates of 99.75% of accuracy, 98.17% of precision, 98.31% of sensitivity and 98.06% of f1-score.

Keywords: GI Tract, WCE, Classification, Transformers, ViT, DeiT.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under Curve*
BERT	*Bidirectional Encoder Representations from Transformers*
BHI	*Biomedical and Health Informatics*
CAD	*Computer Aided Dignosis*
CNN	*Convolutional Neural Network*
DeiT	*Data-efficient image Transformers*
DSSVM	*Deep Sparse Support Vector Machine*
FDA	*Food and Drug Administration*
GPT	*Generative Pre-Training Transformer*
GPU	*Graphics Processing Unit*
HOG	*Histogram of Oriented Gradients*
IA	Inteligência Artificial
IoU	*Intersection over Union*
KNN	*K-Nearest Neighbor*
LBP	*Local Binary Pattern*
MCC	*Matthews Correlation Coefficient*
MLP	*Multilayer Perceptron*
NLP	*Natural Language Processing*
RGB	*Red Green Blue*
RNN	*Recurrent Neural Network*
SDD	*Single Shot MultiBox Detector*
SIFT	*Scale-invariant Feature Transform*
SVM	*Support Vector Machine*

VGG	*Visual Geometry Group*
ViT	*Vision Transformer*
WCE	*Wireless Capsule Endoscopy*

Sumário

1 Introdução

As doenças inflamatórias do trato gastrointestinal apresentam elevada incidência populacional, manifestando altas taxas em países ocidentais, com os maiores índices no norte da Europa, América do Norte, Reino Unido e Austrália (KHORSHIDI et al., 2019). Queixas relacionadas a enfermidades no trato gastrointestinal são frequentes em hospitais, sendo uma das principais causas de internação na população em geral. O efeito dessas doenças pode ser menor se os fatores de risco forem minimizados, tais como consumo excessivo de alimentos industrializados e pouca atividade física, aliados ao diagnóstico precoce e ao tratamento adequado (EWALD et al., 2021).

A Doença de Crohn e a Retocolite Ulcerativa apresentam-se como as principais doenças inflamatórias intestinais, sendo caracterizadas por dor abdominal, febre, sangramento retal, diarreia e graves perdas de peso, além de serem fator de risco para o câncer colorretal (BELÉM; ODA, 2015). Os cânceres relacionados ao trato gastrointestinal, tais como o esofágico, gástrico e colorretal, estão entre os tipos mais comuns em número de novos casos e com altas taxas de mortalidade (SUNG et al., 2021).

A endoscopia é uma das técnicas mais utilizadas para a análise de anormalidades no trato gastrointestinal, entretanto se caracteriza por ser um processo invasivo e doloroso. Em contraste, a obtenção de vídeos por meio de cápsulas endoscópicas sem fio (do inglês *Wireless Capsule Endoscopy* - WCE) configura uma abordagem menos agressiva e desconfortável (ALI et al., 2020). Neste procedimento, o paciente ingere uma pequena cápsula equipada com bateria, fonte de luz, micro-câmera e um emissor de sinal (Figura 1), que percorre passivamente todo o trato gastrointestinal até ser expelida pelo corpo. As imagens capturadas são enviadas para um receptor alojado na cintura do paciente. Por conta do tamanho diminuto da cápsula, esta consegue capturar imagens onde a endoscopia convencional não alcança, entretanto, como percorre todo o trato gastrointestinal, os vídeos geralmente podem atingir até 10 horas de duração. Isto torna a investigação manual do vídeo um procedimento tedioso e propenso a erros, pois depende de concentração contínua durante um longo intervalo de tempo na análise das imagens (IAKOVIDIS et al., 2018). Por esta razão é importante a utilização de alguma automação para ajudar o especialista médico na análise do vídeo.

Técnicas de aprendizagem de máquina, onde um conjunto de dados é utilizado para desenvolver um sistema de diagnóstico assistido por computador (do inglês *Computer Aided Dignosis* - CAD), tem sido utilizadas na análise de imagens médicas desde a década de 1990 (LITJENS et al., 2017). Na última década, as Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* - CNN) se tornaram um modelo de grande sucesso

Figura 1 – Modelos de cápsulas disponíveis comercialmente. Imagens de **(a)** a **(c)** Pillcam, **(d)** MiroCam, **(e)** Olympus e **(f)** Omom.

Fonte: (CIUTI; MENCIASSI; DARIO, 2011).

para o reconhecimento de padrões em imagens.

Para se obter resultados satisfatórios no uso das CNNs, estas devem ser treinadas adequadamente com uma grande, variada e balanceada quantidade de exemplos. Entretanto, no domínio de imagens médicas e, em especial, nas imagens relacionadas ao trato gastrointestinal, há uma pequena quantidade de *datasets* públicos, e estes por sua vez apresentam quantidade limitada de amostras rotuladas (LITJENS et al., 2017).

Arquiteturas baseadas em *Transformer* constituem o estado da arte nas tarefas relacionadas a NLP (do inglês *Natural Language Processing*), onde os modelos BERT (DEVLIN et al., 2018) e GPT-3 (BROWN et al., 2020) figuram entre os mais famosos exemplos. Motivados pelo sucesso de modelos utilizados em NLP, pesquisadores tem evidenciado um interesse crescente em arquiteturas que utilizam mecanismos de atenção para resolver tarefas de visão computacional, provocando o surgimento de propostas de arquiteturas híbridas, mesclando ideias de *Transformers* com CNNs (SHEN et al., 2020), até o surgimento de modelos baseados totalmente em mecanismos de atenção, sem nenhuma camada convolucional (DOSOVITSKIY et al., 2020).

1.1 Objetivos

Diante do contexto apresentado, objetivo principal deste trabalho consiste em estruturar um método computacional para análise de imagens provenientes de vídeos de endoscopia por cápsula (WCE), que utilize arquitetura baseada em *Transformers*, para a indução de um modelo que seja capaz de classificar imagens do trato gastrointestinal como normal ou contendo alguma anormalidade, visando auxiliar o especialista médico com um procedimento que provê uma segunda opinião e proporciona maior eficiência no diagnóstico.

1.1.1 Objetivos Específicos

Especificamente, este trabalho busca os seguintes objetivos aplicados ao problema de classificação de imagens do trato gastrointestinal obtidas por exames de endoscopia por cápsula:

- Analisar e desenvolver arquiteturas de redes neurais utilizando *Transformers* para classificação de imagens obtidas por WCEs;
- Analisar os resultados obtidos pelo modelo proposto utilizando o *dataset* público Kvasir-Capsule;
- Comparar o desempenho do método proposto com trabalhos encontrados na literatura;
- Analisar as vantagens e limitações do método proposto.

1.1.2 Contribuições

Destacam-se como principais contribuições:

- A utilização de redes baseadas em *Transformers* para a classificação de imagens do trato gastrointestinal;
- Desenvolvimento de um método automatizado capaz de detectar anomalias em imagens de exames de endoscopia por cápsula.

1.2 Organização do Trabalho

Este trabalho está estruturado da seguinte forma:

- O Capítulo 2 descreve trabalhos relacionados à análise de imagens médicas específicas do trato gastrointestinal, que utilizam diversas soluções baseadas em *deep learning* visando detectar diferentes tipos de anomalias;
- O Capítulo 3 trata da fundamentação teórica no qual são abordados conceitos refentes aos exames de imagens por cápsulas e arquiteturas *Transformer*;
- O Capítulo 4 apresenta as etapas adotadas que compõem a metodologia proposta para a classificação de imagens do trato gastrointestinal, as técnicas utilizadas nos processos de treinamento e teste dos modelos, bem como na avaliação de resultados;
- O Capítulo 5 trata sobre os resultados obtidos e discussões em relação aos experimentos realizados utilizando o método proposto, assim como apresenta comparação com outros trabalhos encontrados na literatura;
- O Capítulo 6 apresenta as considerações finais sobre os resultados, propostas de trabalhos futuros e artigo científico publicado.

2 Trabalhos Relacionados

Neste capítulo serão apresentados trabalhos publicados que abordam técnicas de aprendizagem de máquina para a detecção de doenças do trato gastrointestinal em vídeos obtidos por cápsulas endoscópicas.

Devido a facilidade de uso e baixo desconforto provocado ao paciente, WCEs tem se constituído no exame mais indicado para investigar anomalias presentes no intestino delgado, tais como sangramento intestinal, úlceras, doença de Crohn, doença celíaca, pólipos e tumores (KRÖNER et al., 2021).

Diversos trabalhos abordaram a detecção de anomalias no trato gastrointestinal analisando imagens endoscópicas utilizando diferentes enfoques, desde a utilização de técnicas tradicionais para extração de características e em conjunto com a utilização de classificadores, até o uso de técnicas avançadas de *deep learning* como a aplicação de modelos populares de CNNs e a utilização de redes *Transformer*.

Dentre os trabalhos que utilizaram técnicas tradicionais para detecção de anomalias em imagens obtidas por WCE, destaca-se o trabalho de Nawarathna et al. (2014), onde foi utilizada a técnica de *Local Binary Pattern* (LBP) para a extração de características de textura da imagem para uso como entrada de um algoritmo classificador *K-Nearest Neighbor* (KNN), tendo como enfoque principal a distinção de imagens saudáveis das com alguma anormalidade na mucosa. Foi utilizado um *dataset* próprio composto por 5 vídeos contendo 1250 imagens rotuladas em 5 classes, obtendo ao final 92,00% de sensibilidade e 91,80% de especificidade.

O trabalho de Cong et al. (2015) propôs nova variante do SVM denominada *Deep Sparse Support Vector Machine* (DSSVM). Em seu método, as imagens de entrada são quebradas em *super-pixels*, de onde são extraídas características de cor e textura, para serem classificadas pelo DSSVM, que seleciona as caracterísicas mais relevantes ao problema. Dessa forma, o modelo atinge boa acurácia, enquanto reduz a complexidade do cálculo. Em seu melhor resultado alcançou 0,9079 de *Area Under Curve* (AUC).

Yuan, Li e Meng (2016) propuseram um método para classificação de pólipos em imagens de WCEs. Foram utilizadas as técnicas *Scale-invariant Feature Transform* (SIFT) e LBP para extração de características em conjunto com o classificador *Support Vector Machine* (SVM). Em seus experimentos foi utilizado *dataset* privado cedido pelo Qilu Hospital da China, composto por 20 WCEs, onde foram extraídos 2500 imagens, sendo 2000 normal e 500 contendo pólipos. Como melhor resultado atingiram 93,20% de acurácia, 90,88% de sensibilidade e 94,54% de especificidade. Apesar de seu resultado superior, este trabalho focou apenas na classificação em duas classes.

Com o passar do tempo, técnicas tradicionais de extração de características foram sendo substituídas por métodos baseados em *deep learning* para as tarefas de classificação de imagens médicas. Jia e Meng (2017) adotaram um abordagem híbrida, com o uso de técnicas tradicionais em conjunto com arquitetura própria de CNN para extração de características, utilizando ao final uma camada classificadora com o objetivo de detectar sangramentos em imagens obtidas por WCEs. Este trabalho utilizou *dataset* próprio composto por 1500 imagens, das quais 300 apresentavam algum tipo de sangramento. Como resultado atingiram 91,00% para sensibilidade, 94,69% para precisão e 92,85% para *f1-score*.

Em virtude da baixa quantidade de *datasets* públicos de imagens médicas, geralmente associados a poucas amostras rotuladas, o uso das técnicas de transferência de aprendizagem e aumento de dados se tornou comum em trabalhos que analisam imagens médicas (LITJENS et al., 2017). No trabalho de Li et al. (2017) foi realizado estudo comparativo utilizando várias arquiteturas CNNs diferentes, tais como LeNet, AlexNet, GoogLeNet, e VGG, visando a detecção de sangramentos ou hemorragias gastrointestinais. Este trabalho adotou técnicas de transferência de aprendizado, assim como diversas técnicas de aumento dados, como espelhamento, alterações de luminosidade e ruído, com o intuito de melhorar o desempenho dos modelos treinados. Neste estudo, VGG se mostrou o modelo mais eficiente, alcançando taxas de 99,10% para sensibilidade, 98,65% para precisão e 98,87% para *f1-score*. Após vários experimentos, demonstraram a melhora nos resultados com o uso das técnicas de aumento de dados e transferência de aprendizado. Entretanto, o desempenho ao analisar imagens com baixo contraste foi ligeiramente inferior, principalmente em imagens com bolhas, onde foram falsamente classificadas como hemorragia. Klang et al. (2020) propuseram um algoritmo para detecção automatizada de úlceras em imagens captadas por WCE. Foi utilizado modelo CNN Xception pré-treinado com os pesos do ImageNet, aplicado em um *dataset* privado composto de 17640 imagens de 49 pacientes. Por sua vez, o trabalho de Saito et al. (2020) focou na detecção de pólipos, tumores e demais lesões semelhantes. Foi utilizado *dataset* privado, composto de vídeos obtidos por WCEs de hospitais do Japão, totalizando 30584 imagens de exames de 292 pacientes. Foi proposta arquitetura CNN baseada em *Single Shot MultiBox Detector (SSD)*, onde obteve resultados satisfatórios de acordo com cada subtipo de lesão (pólipos, nódulos, tumores epiteliais, tumores submucosos e estruturas venosas), atingindo taxas de sensibilidade de 90,70% e especificidade de 79,80%.

Nos modelos CNN, as operações de convolução são utilizadas para agregar características espaciais locais, entretanto falham em capturar dependências globais. Visando contornar este problema, e devido ao sucesso dos mecanismos de atenção utilizados nas redes *Transformer* em tarefas de NLP, pesquisadores passaram a utilizar as ideias de tais redes em tarefas de visão computacional. Modelos *Transformer* tornaram-se uma solução atraente devido à sua capacidade de codificar dependências globais de longo alcance e

aprender representações de características altamente eficazes (SHAMSHAD et al., 2022).

No trabalho de Muruganantham e Balakrishnan (2022) foi apresentada arquitetura híbrida para análise de imagens obtidas por WCE. O processo se inicia com a utilização de camadas compostas por mecanismos de atenção para estimar mapas de atenção de baixo e alto nível, com o objetivo de destacar as áreas de lesões nas imagens analisadas. Após esta etapa, os mapas de atenção estimados são fundidos com as imagens WCE originais para realçar a região da lesão presente nas imagens de entrada, sendo então repassada para camadas convolucionais para a etapa de classificação. Em seus experimentos foram utilizados dois *datasets* públicos, onde seus melhores resultados foram alcançados com o Kvasir-Capsule (SMEDSRUD et al., 2021), utilizando quantidades iguais de exemplos para 3 classes analisadas, obtendo ao final 95,36% de acurácia, 95,20% de precisão e 95,25% de *f1-score*.

Abordagem semelhante foi adotada por Srivastava et al. (2022). Foi proposta arquitetura híbrida denominada FocalConvNet, composta por blocos com mecanismos de atenção e camadas convolucionais. Em seus experimentos foram utilizadas 47153 imagens do *dataset* Kvasir-Capsule, sendo adotada uma proporção incomum em sua divisão, utilizando 49% do total de imagens para treino e o restante para teste. Ao final alcançaram resultado modesto, com 63,73% para acurácia, 75,57% para precisão, 63,73% para sensibilidade e 67,34% para *f1-score*.

Dosovitskiy et al. (2020) apresentaram em seu trabalho o modelo ViT, que é uma arquitetura derivada dos modelos utilizados em NLP, mas aplicada à classificação de imagens, utilizando *patches* de imagens como entrada. Este trabalho apresentou excelentes resultados com o uso de modelo baseado em *Transformer* treinado em um grande *dataset* privado de imagens rotuladas (JFT-300M, 300 milhões de imagens). No entanto, seus autores destacam que o ViT precisa de grandes *datasets* para obter resultados comparáveis às CNNs, e o treinamento desses modelos envolve extensos recursos computacionais. Visando contornar a necessidade de *datasets* muito grandes para treinar adequadamente modelos baseados em *Transformer*, Bai et al. (2022) propuseram arquitetura híbrida, utilizando o modelo ViT juntamente com camadas convolucionais de *pooling* após as camadas do *Transformer*, com o objetivo de reduzir a dimensão dos mapas de características, diminuindo dessa forma a quantidade de parâmetros e informações redundantes na rede. Os experimentos foram executados utilizando o *dataset* público Kvasir-Capsule, sendo o modelo treinado totalmente neste *dataset*, empregando todas as 11 classes da categoria *luminal findings*. Em comparação com diversos modelos CNN, seus resultados se mostraram superiores, alcançando taxas de acurácia de 79,15%.

Os trabalhos abordados neste capítulo estão resumidos na Tabela 1. Estes representam amostra da evolução das técnicas aplicadas para a detecção de anomalias em imagens obtidas por WCE. Apesar de alguns trabalhos demonstrarem métodos utilizando arquitetu-

ras híbridas, agrupando características de CNNs e *Transformers*, nenhum trabalho aplicou arquitetura pura de *Transformer*, sem qualquer camada convolucional, para a classificação de imagens WCE. Neste trabalho foi desenvolvido um método para a classificação de anormalidades do trato gastrointestinal obtidas por imagens de exames de endoscopia por cápsula, com a utilização de redes neurais baseadas totalmente em *Transformers*, com uso de técnicas de transferência de aprendizado e aumento de dados. A fundamentação teórica utilizada para o desenvolvimento do método, bem como o método em si são assuntos dos próximos capítulos desta dissertação.

Tabela 1 – Lista de trabalhos relacionados destacados na revisão da literatura. Resultados exibidos para as métricas acurácia (ACU), sensibilidade (SEN), especificidade (ESP), precisão (PRE), *f1-score* (F1) e AUC.

Autores	**Método**	**Dataset (Amostras)**	**Resultado**
Nawarathna et al. (2014)	LBP + KNN	Privado (1250)	92,00% de SEN 91,80% de ESP
Cong et al. (2015)	Histrogramas de Cor + LBP + DSSVM	Privado (3800)	0,9079 de AUC
Yuan, Li e Meng (2016)	SIFT + LBP + SVM	Privado (2500)	93,20% de ACU 90,88% de SEN 94,58% de ESP
Jia e Meng (2017)	CNN	Privado (1500)	94,69% de PRE 91,00% de SEN 92,85% de F1
Li et al. (2017)	CNN	Privado (12090)	98,65% de PRE 99,10% de SEN 98,87% de F1
Klang et al. (2020)	CNN	Privado (17640)	96,70% de ACU 96,80% de SEN 96,60%de ESP
Saito et al. (2020)	CNN	Privado (30584)	90,70% de SEN 79,80% de ESP
Muruganantham e Balakrishnan (2022)	CNN + Mecanismo de atenção	Kvasir-Capsule (2400)	95,36% de ACU 95,20% de PRE 95,25% de F1
Srivastava et al. (2022)	CNN + Mecanismo de atenção	Kvasir-Capsule (47153)	63,73% de ACU 75,57% de PRE 67,34% de F1
Bai et al. (2022)	CNN + *Transformer*	Kvasir-Capsule (41510)	79,15% de AUC

3 Fundamentação Teórica

Este capítulo apresenta os conceitos explorados para o desenvolvimento da pesquisa que resultou em um método automatizado de diagnóstico de anormalidades no trato gastrointestinal em imagens obtidas por cápsulas endoscópicas.

3.1 Cápsulas Endoscópicas sem Fio

Cápsulas endoscópicas sem fio (WCE) são dispositivos revolucionários para a realização de exames de forma não invasiva e indolor do trato gastrointestinal. Esta tecnologia foi apresentada em 2000 pela empresa israelense Given Imaging e recebeu aprovação da *Food and Drug Administration* - FDA dos EUA em 2001. WCE é um dispositivo em forma de cápsula com dimensão aproximada de 26mm × 11mm, composto por cúpula óptica, fontes de luz, sensor de imagem, bateria e um sistema transmissor de rádio, capaz de fornecer a visualização interna de todo o trato gastrointestinal (JIA et al., 2020). Na Tabela 2 são listados principais modelos de cápsulas vendidas comercialmente.

Tabela 2 – Especificações dos principais modelos de WCE vendidos comercialmente.

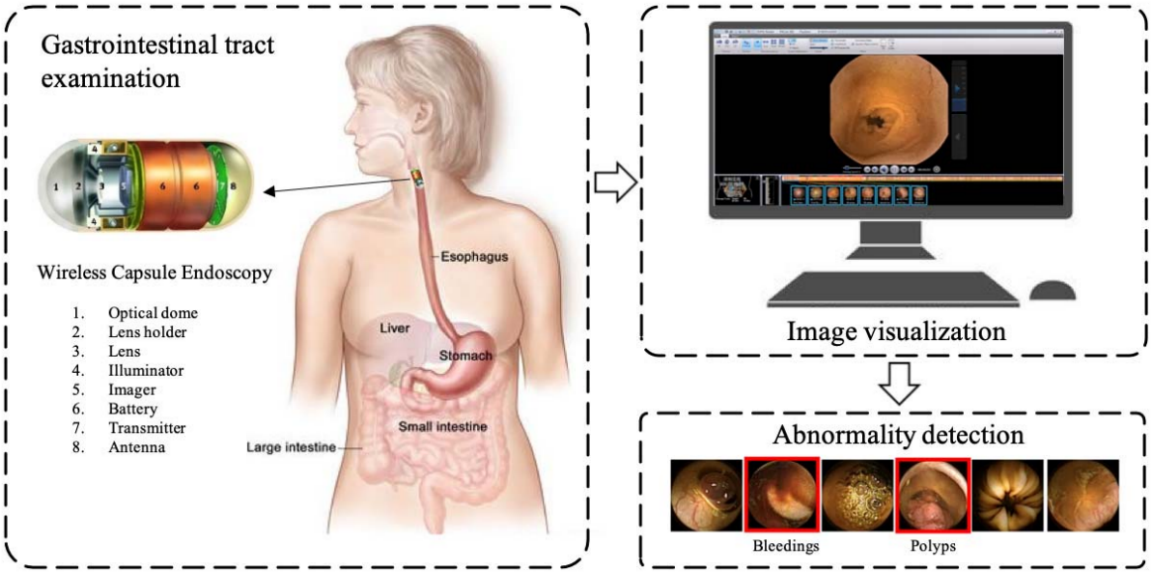
Modelo	**Fabricante**	**Tamanho (mm x mm)**	**Duração da bateria**
PillCam SB	Given Imaging Inc. (Israel)	26 x 11	8h-9h
EndoCapsule	Olympus Inc. (Japão)	26 x 11	8h-10h
MiroCam	Intromedic Inc. (Coréia do Sul)	24 x 11	11h
OMOM	Chongqing Jinshan Science & Technology (China)	24 x 11	7h-9h
PillCam ESO	Given Imaging Inc. (Israel)	26 x 11	20min
PillCam COLON	Given Imaging Inc. (Israel)	31 x 11	10h

Fonte: adaptado de (CIUTI; MENCIASSI; DARIO, 2011).

Após ser ingerida pelo paciente, a cápsula percorre o corpo humano ao longo de todo sistema digestivo, passando pelo esôfago, estômago, piloro, duodeno, intestino delgado e cólon, até ser excretada pelo ânus. Ao longo do percurso, são registradas duas ou mais imagens coloridas do trato gastrointestinal a cada segundo, por cerca de 8 horas de exame ou até que as baterias do dispositivo se esgotem. Essas imagens são então comprimidas

e transmitidas sem fio para um dispositivo de gravação de dados que geralmente fica preso à cintura do paciente. Ao final do processo, cerca de 50.000 imagens são geradas e transferidas para um computador (JIA et al., 2020). Este procedimento está retratado na Figura 2. Em comparação com as modalidades convencionais de endoscopia, em que muitos pacientes consideram a preparação do exame um procedimento desagradável em relação à percepção de desconforto, constrangimento e risco, a utilização de WCE se caracteriza por ser um procedimento mais simples e agradável, sem a necessidade de sedação por parte do paciente (ELIAKIM et al., 2006).

Figura 2 – Fluxo para o diagnóstico de anormalidades no trato gastrointestinal com WCE.



Fonte: (JIA et al., 2020).

O exame por cápsula endoscópica é um procedimento relativamente simples para o paciente. Após a ingestão da cápsula, o paciente pode continuar suas atividades diárias normais enquanto a pílula atravessa todo o trato gastrointestinal. Segundo Wang et al. (2013), o consumo de líquidos e alimentação leve pode ocorrer algumas horas após a ingestão da pílula. Comumente, os pacientes são instruídos a observar a passagem da cápsula em seus movimentos intestinais, ou são solicitados a fazer uma radiografia abdominal se não for observada a entrada da cápsula no cólon durante a revisão do exame. De uma forma geral, espera-se que todo o intestino delgado possa ser visualizado durante a vida útil padrão de 8 horas da bateria. No entanto, fatores como esvaziamento gástrico lento, detritos no intestino ou trânsito intestinal podem impedir um exame completo em 17% a 25% dos casos, levando à necessidade de preparo intestinal adicional em alguns pacientes (RONDONOTTI et al., 2005).

Embora WCE tenha mostrado vantagens significativas sobre as endoscopias tradicionais para inspecionar o trato gastrointestinal, ainda há espaço para melhorias. Um problema associado a esta nova tecnologia é que as sequências de imagens produzidos por

WCEs são revisadas manualmente, sendo uma tarefa trabalhosa e demorada (GOSSUM et al., 2009). De acordo com Hwang (2011), o tempo médio de leitura de um vídeo produzido por sequência de imagens WCE é de 45 minutos, podendo variar de 30 a 75 minutos. Destaca ainda que as imagens com alguma anormalidade ocupam menos de 5% do total de imagens coletadas. Nesse sentido, o desenvolvimento de sistemas de diagnóstico assistido por computador para análise automatizada de imagens WCE é altamente desejável.

Cada segmento do trato gastrointestinal é caracterizado por diferentes propriedades anatômicas e fisiológicas, levando a diferentes desafios para o projeto de cápsulas endoscópicas. Dessa forma, a realização de exames de esôfago, intestino delgado e cólon necessitam de diferentes preparações e modelos de cápsulas. A seguir as principais características destas modalidades de exames.

3.1.1 Exames de Esôfago

As cápsulas PillCam ESO são os modelos indicados para exames esofágicos. Conforme pode ser conferido na Tabela 2, suas dimensões são semelhantes ao modelo PillCam SB, entretanto a duração da bateria é de apenas 20 minutos, em contraste à autonomia dos outros modelos que variam de 8 a 11 horas. Outra diferença é que as cápsulas PillCam ESO possuem duas câmeras, localizadas em ambas as extremidades, gravando imagens a uma taxa de 18 quadros por segundo. Conforme destacam Ciuti, Menciassi e Dario (2011), devido aos movimentos peristálticos e à rápida passagem do alimento pelo esôfago, as cápsulas endoscópicas projetadas para esta região requerem alta taxa de quadros por segundo e bateria de baixa duração. Como receptores de sinal, são utilizados sensores colados no tórax do paciente.

A preparação deste exame é mais simples comparada aos exames do intestino delgado e cólon, sendo necessário apenas que o paciente esteja em jejum por pelo menos 2 horas. Por conta da rápida realização do exame, e de movimentos específicos que o paciente deve realizar, o procedimento é executado todo em ambulatório médico. O procedimento se inicia com o paciente ingerindo a cápsula juntamente com 100 ml de água, em seguida é colocado em decúbito dorsal, ou seja, deitado de costas. Após 2 minutos o paciente eleva o tórax a 30º, permanecendo nesta posição por 1 minuto, para em seguida elevar novamente o tórax para 60º e permanece nesta posição por mais 1 minuto. Por fim, o paciente fica em pé, por aproximadamente 15 minutos para maximizar o tempo para a cápsula capturar imagens enquanto atravessa o esôfago (WANG et al., 2013).

3.1.2 Exames de Intestino Delgado

Os exames mais indicados para o uso de WCE são os relacionados ao intestino delgado, concentrando dessa forma a atenção dos fabricantes. Os modelos de cápsula

utilizados na investigação do intestino delgado são PillCam SB, EndoCapsule, MiroCam e OMOM, com a capacidade de duração da bateria variando de 8 a 11 horas (CIUTI; MENCIASSI; DARIO, 2011). Entre as principais anormalidades encontradas nesta região, destacam-se a doença de Crohn, a hemorragia digestiva obscura, a doença celíaca, os pólipos e tumores (KWACK; LIM, 2016).

Ao contrário dos instrumentos utilizados na endoscopia convencional, a intensidade da iluminação fornecida por WCE não varia conforme a necessidade. Além disso, uma lesão não pode ser lavada ou examinada repetidamente. Em alguns pacientes, a cápsula endoscópica não consegue visualizar todo o intestino delgado devido ao conteúdo intestinal. Durante o exame, é possível notar o escurecimento progressivo das imagens a medida que a cápsula se move, provavelmente devido à presença de bile e alimentos não absorvidos. Em pacientes que executaram preparo específico para colonoscopia, as imagens são significativamente mais claras e nítidas, com melhor qualidade. Por conta disto, o preparo para os exames de intestino delgado é mais complexo, exigindo uma dieta mais rigorosa (DAI et al., 2005).

Como sistema receptor de sinal, são utilizados sensores na parede abdominal, conectados ao gravador de dados usado pelo paciente. Após a ingestão da cápsula, os pacientes são instruídos a manter um registro dos sintomas e monitorar as luzes do gravador de dados para confirmar que o sinal está sendo recebido. Os pacientes são encorajados a evitar exercícios ou atividades que possam fazer com que os sensores se soltem. Uma dieta composta por líquidos claros é permitida após 2 horas e por refeição leve após 4 horas. O sistema de registro de dados reutilizável pode ser desconectado do paciente após o término da vida útil da bateria. A cápsula é descartável e projetada para ser excretada. O gravador de dados é posteriormente conectado a uma estação de trabalho para transferência das imagens registradas (WANG et al., 2013).

3.1.3 Exames do Cólon

O exame de cólon por cápsula endoscópica pode ser uma boa alternativa em pacientes que recusam a colonoscopia convencional ou quando a colonoscopia convencional é inadequada ou impossível. No trabalho de Gossum et al. (2009), foi realizada análise de 328 casos de utilização de WCE para exames de cólon, obtendo 92% de taxa de visualização completa do cólon antes do fim da vida útil da bateria. Embora a utilização de WCE tenha mostrado capacidades de detecção semelhantes quando comparada a exames com colonoscopia convencional, esta permanece mais precisa, aliada ao fato de que permite a remoção simultânea de pólipos no momento da realização do exame. O dispositivo WCE indicado para este tipo de exame é o PillCam COLON e a preparação é semelhante à realizada no exame de intestino delgado (WANG et al., 2013).

3.2 Transformers

No trabalho de Vaswani et al. (2017) foi proposta a arquitetura clássica da rede *Transformer*, projetada para tarefas de tradução em NLP, sendo composta apenas por mecanismos de atenção, dispensando o uso de recorrência ou camadas convolucionais. Ao contrário das Redes Neurais Recorrentes (do inglês *Recurrent Neural Network - RNN*), que processam elementos de uma sequência de forma recursiva e só aprendem contextos de curto prazo, os *Transformers* podem processar sequências completas em paralelo, aprendendo assim relacionamentos de longo alcance (KHAN et al., 2021).

Existem duas ideias chave que contribuíram para o desenvolvimento de modelos *Transformer*. A primeira é o mecanismo de auto-atenção (do inglês *self-attention*), que permite capturar dependências globais entre elementos de uma mesma sequência. A outra ideia chave é o mecanismo de atenção multi-cabeça (do inglês *multi-head attention*), que possibilita o processamento paralelo de múltiplos blocos de atenção e permite que o modelo lide conjuntamente com informações de diferentes subespaços de representação em diferentes posições (VASWANI et al., 2017).

Conforme mostrado na Figura 3, a arquitetura da rede *Transformer* é composta por dois ramos, o codificador (*encoder*) e o decodificador (*decoder*), contendo múltiplos blocos *Transformer* de mesma arquitetura. O codificador recebe uma representação da sequência de entrada, gerando informações sobre quais partes são relevantes entre si, enquanto o decodificador usa os dados do codificador para gerar as sequências de saída ao incorporar informações contextuais. Cada bloco *Transformer* é constituído por um mecanismo de atenção multi-cabeça, uma rede neural *feed-forward*, conexões residuais e camadas de normalização (HAN et al., 2022).

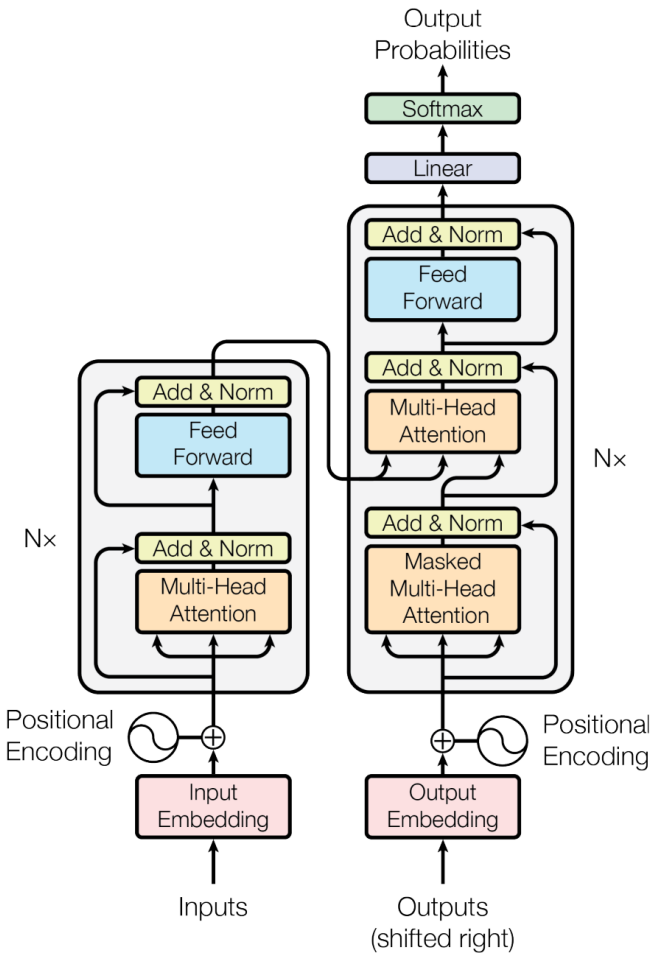
3.2.1 Mecanismo de Auto-Atenção

O sucesso dos modelos *Transformer* é atribuído ao mecanismo de auto-atenção devido à sua capacidade de modelar dependências de longo alcance. A ideia chave por trás do mecanismo auto-atenção é aprender o auto-alinhamento, ou seja, determinar a importância relativa de um único *token*[1] em relação a todos os outros *tokens* em uma sequência (BAHDANAU; CHO; BENGIO, 2014).

Dada uma sequência de itens, o mecanismo de auto-atenção estima a relevância de um item em relação aos outros itens da sequência, como, por exemplo, quais palavras provavelmente virão juntas em uma frase. O mecanismo de auto-atenção é um componente integral da rede *Transformer*, que modela explicitamente as interações entre todas as entidades de uma sequência para tarefas de predição. Basicamente, uma camada de auto-

[1] Nome que se dá a cada item de uma frase ou sequência de palavras a ser analisada em NLP.

Figura 3 – Arquitetura do modelo *Transformer*.



Fonte: (VASWANI et al., 2017).

atenção atualiza cada componente de uma sequência agregando informações globais da sequência de entrada completa (KHAN et al., 2021).

Conforme Han et al. (2022) explicam em seu trabalho, o funcionamento do mecanismo de auto-atenção ocorre da seguinte forma: o vetor de entrada é transformado em três diferentes vetores, o vetor de pesquisa q, vetor de chaves k, e o vetor de valores v, todos contendo a mesma dimensão d. Em seguida, os vetores derivados de diferentes entradas são então agrupados em três matrizes diferentes, chamadas de Q, K e V. Ao final, a função de atenção entre os diferentes vetores de entrada é calculada conforme a sequência de passos abaixo:

- **Passo 1**: Calcula as pontuações entre os diferentes vetores de entrada;

$$S = Q \cdot K^T \tag{3.1}$$

- **Passo 2**: Normaliza as pontuações para estabilidade do gradiente;

$$S_n = S/\sqrt{d_k} \tag{3.2}$$

- **Passo 3**: Calcula as probabilidades;

$$P = softmax(S_n) \tag{3.3}$$

- **Passo 4**: Gera a matriz de atenção.

$$Z = V \cdot P \tag{3.4}$$

Todo este processo pode ser simplificado na Equação 3.5.

$$Atenção(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{3.5}$$

Resumidamente, a lógica por trás da Equação 3.5 pode ser explicada da seguinte forma: são calculadas pontuações entre cada par de vetores diferentes, onde essas pontuações determinam o grau de atenção em relação às outras palavras ao codificar a palavra na posição atual. Ato contínuo as pontuações são normalizadas para melhorar a estabilidade do gradiente e aperfeiçoar o treinamento, em seguida são calculadas probabilidades para as pontuações, e por fim, cada vetor de valor é multiplicado pela soma das probabilidades. Dessa forma, vetores com maiores probabilidades recebem foco adicional nas camadas seguintes.

Como a arquitetura *Transformer* não utiliza recorrência nem convolução, para que o modelo entenda a ordem da sequência de entrada devem ser fornecidas informações sobre a posição relativa ou absoluta das palavras na referida sequência. Esta informação é fornecida pelo *positional encoding*, que é um vetor com dimensão d, obtido pelo cálculo de seno e cosseno, dependendo da posição da palavra na sequência de entrada (VASWANI et al., 2017). Dessa forma, cada elemento do *positional encoding* corresponde a uma senoide e permite que o modelo *Transformer* aprenda a tratar posições relativas e extrapolar para comprimentos de sequência mais longos durante a inferência (HAN et al., 2022).

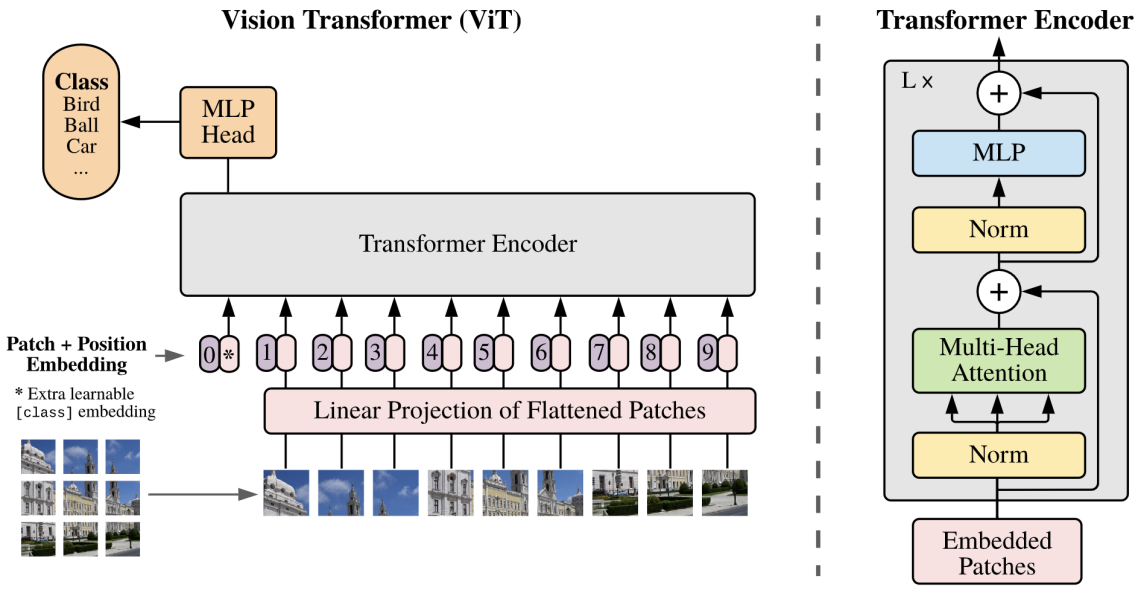
3.2.2 Mecanismo de Atenção Multi-Cabeça

Um único mecanismo de atenção limita a habilidade do modelo em focar em uma ou mais posições específicas sem influenciar a atenção em outras posições igualmente importantes ao mesmo tempo. Utilizar outros mecanismos de atenção em paralelo permite ao modelo lidar conjuntamente com informações de diferentes subespaços de representação em diferentes posições, dessa forma, mecanismos de atenção multi-cabeça são utilizados para aumentar o desempenho da camada de auto-atenção. Especificamente, diferentes matrizes de consulta, chave e valor são usadas para cabeças diferentes, e essas matrizes podem projetar vetores de entrada em diferentes subespaços de representação (HAN et al., 2022).

3.2.3 Vision Transformer (ViT)

Motivados pelo sucesso alcançado por *Transformers* em tarefas relacionadas a NLP, Dosovitskiy et al. (2020) propuseram o modelo *Vision Transformer* (ViT) aplicado a tarefa de classificação de imagens. Sua ideia foi utilizar a arquitetura original do modelo *Transformer* com as menores modificações possíveis. Seus autores utilizaram abordagem semelhante a tarefas de processamento de linguagem natural, onde uma sequência de palavras é quebrada em vários *tokens*, sendo em seguida gerado o *embedding* destes *tokens* para ser alimentado o *encoder* do *Transformer*. Para o processamento das imagens, o modelo ViT divide a imagem a ser analisada em vários *patches*, para em seguida ser gerada uma sequência linear de *embeddings*, que será utilizada como entrada para o *Transformer*. Portanto, os *patches* da imagem são tratados da mesma forma que os *tokens* em um modelo utilizado em NLP. Na Figura 4 é apresentada a visão geral deste modelo.

Figura 4 – Arquitetura do modelo ViT.



Fonte: (DOSOVITSKIY et al., 2020).

Para lidar com imagens 2D, a imagem de entrada $X \in \mathbb{R}^{h \times w \times c}$ é remodelada em uma sequência de *patches* $X_p \in \mathbb{R}^{n \times (p^2 \cdot c)}$, sendo (h, w) a resolução original da imagem, c o número de canais, (p, p) a resolução de cada *patch*, e $n = h \cdot w/p^2$ o número total de *patches*, que também corresponde ao tamanho da sequência de entrada do *Transformer*. Em seguida, é criada projeção linear da sequência de *patches* mapeada em vetores de dimensão constante D. Todos os vetores do *Transformer* são de mesma dimensão. A projeção linear treinável resultante é chamada de *patch embedding* (DOSOVITSKIY et al., 2020).

Ao *patch embedding* são adicionados ainda o *positional encoding*, com o objetivo de reter informações posicionais dos *patches*, e o *class token*, que é aprendido pela rede e

serve como uma representação da imagem de entrada. Em seguida os dados são enviados como entrada para o codificador *Transformer*, cuja saída alimenta camada MLP (do inglês *Multilayer Perceptron*), onde é realizada a classificação final. Comparado ao modelo proposto por Vaswani et al. (2017) para tarefas de NLP, o modelo ViT utiliza apenas o ramo do codificador para a tarefa de classificação de imagens, onde este é composto por vários blocos *Transformer*, variando de 12 a 32 blocos conforme a versão do modelo (Tabela 3). Cada bloco *Transformer* é composto por camadas alternadas de mecanismos de atenção multi-cabeça e MLP, precedidas por camadas de normalização e sucedidas por conexões residuais, conforme consta no detalhamento do codificador *Transformer* na Figura 4.

Tabela 3 – Detalhamento das variações do modelo ViT.

Modelo	**Blocos**	**Cabeças**	**Parâmetros**
ViT-Base (ViT-B)	12	12	86 milhões
ViT-Large (ViT-L)	24	16	307 milhões
ViT-Huge (ViT-H)	32	16	632 milhões

O ViT produz resultados modestos quando treinado em conjuntos de dados de tamanho médio, tais como o ImageNet, alcançando acurácia alguns pontos percentuais abaixo de modelos ResNet de tamanho compatível. No entanto, Dosovitskiy et al. (2020) destacam em seu trabalho que, quando o treinamento é realizado utilizando um grande *dataset*, como JFT-300M (SUN et al., 2017) com 300 milhões de imagens, o modelo ViT se aproximou e até mesmo superou o estado da arte em vários desafios de reconhecimento de imagem. Em função disto, o uso padrão do modelo ViT é fazer um pré-treinamento em um *dataset* grande, em seguida realizar ajuste fino no *dataset* desejado. Este é um passo importante, pois *Transformers* não generalizam bem quando treinados com quantidade insuficiente de dados, devido a falta de vieses indutivos inerentes às CNNs, tais como a equivariância de translação e localidade (HAN et al., 2022).

De uma forma geral, o funcionamento do modelo ViT pode ser sintetizado conforme a ordem abaixo:

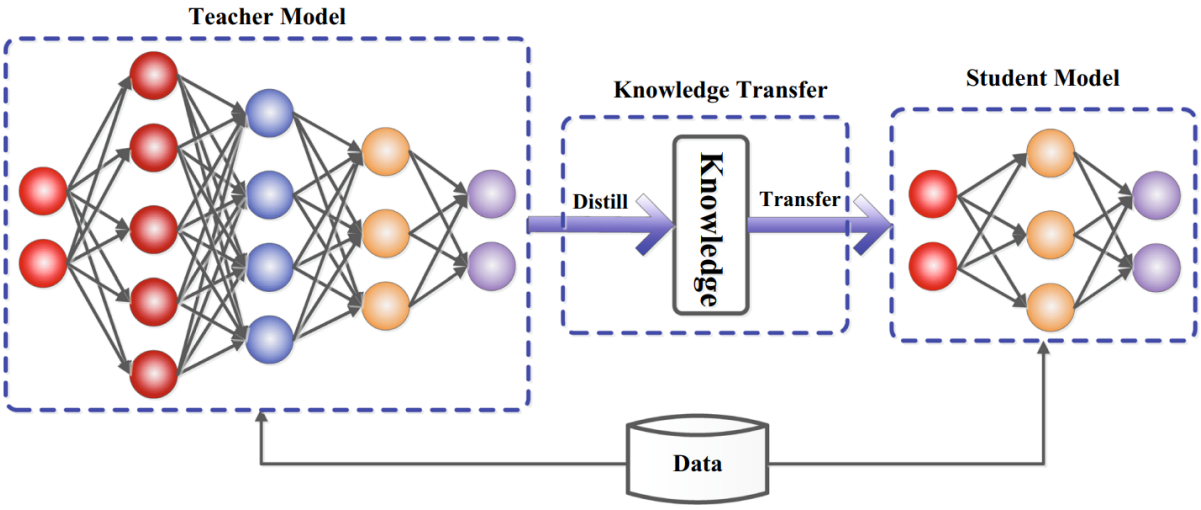
- Divide a imagem de entrada em *patches* de tamanho fixo;
- Cria a sequência de *patches*;
- Calcula as projeções do *patch embedding*;
- Adiciona o *positional encoding* e o *class token* ao *patch embedding*;
- Envia sequência ao codificador *Transformer*;
- Realiza o pré-treinamento do modelo em um conjunto grande de imagens;
- Executa o ajuste fino na base de dados alvo da classificação.

3.2.4 Data-efficient image Transformers (DeiT)

Visando resolver o problema da necessidade de grandes *datasets* para o pré-treinamento do modelo ViT, Touvron et al. (2021) apresentaram o modelo *Data-efficient image Transformers* (DeiT), propondo uma arquitetura baseada em *Transformers* sem nenhuma operação de convolução, tal qual o modelo ViT, alcançando resultados competitivos, utilizando na etapa de pré-treinamento um *dataset* bem menor graças a um novo procedimento de destilação, empregando uma estratégia professor-aluno específica para *Transformers*, bem como o uso de forte conjunto de técnicas de regularização e aumento de dados.

Destilação de conhecimento (do inglês *knowledge distillation*) é uma técnica de compressão de modelo na qual um modelo pequeno (aluno) é treinado para imitar um modelo maior (professor) pré-treinado. Esta técnica foi originalmente proposta por Buciluǎ, Caruana e Niculescu-Mizil (2006) e depois aperfeiçoada por Hinton et al. (2015). Na destilação, o conhecimento é transferido do modelo professor para o aluno minimizando uma função de perda na qual o alvo é a distribuição de probabilidades de classe prevista pelo modelo professor. A ideia principal é que o modelo aluno mimetize o modelo professor para obter um desempenho competitivo ou mesmo superior (GOU et al., 2021). Na Figura 5 é apresentada a arquitetura padrão do modelo professor-aluno utilizando destilação de conhecimento.

Figura 5 – Arquitetura do modelo professor-aluno com destilação de conhecimento.



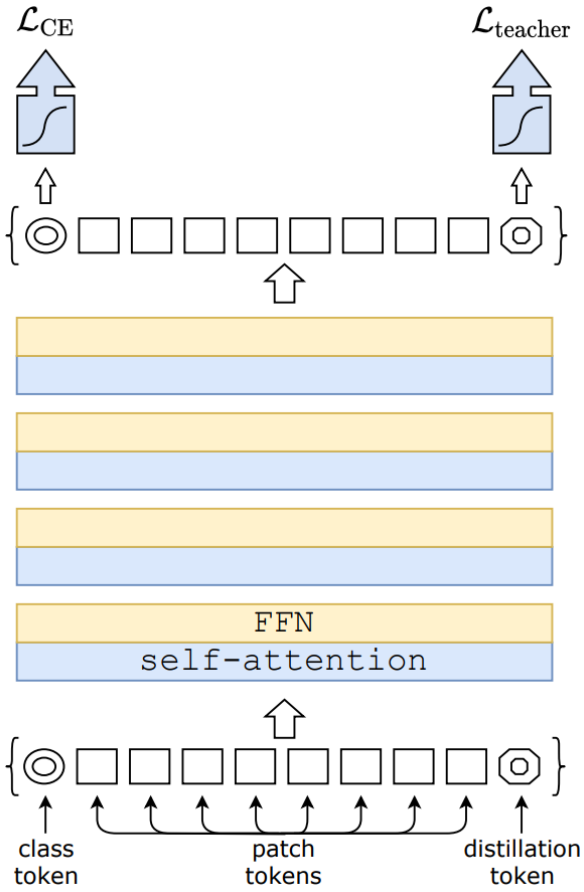
Fonte: (GOU et al., 2021).

A arquitetura do modelo DeiT proposta por Touvron et al. (2021) é bastante similar ao modelo ViT tratado anteriormente, onde a única diferença consiste na adição do *distillation token*, juntamente com o *class token* e o conjunto de *patch tokens*, conforme destacado na Figura 6. O *distillation token* é usado de forma semelhante ao *class token*, onde estes interagem com os *patch tokens* por meio do mecanismo de auto-atenção. Dessa

forma, permite que o modelo aprenda também com a saída do professor, como em uma destilação regular. O objetivo principal do *distillation token* é reproduzir o rótulo previsto pelo professor, enquanto o objetivo do *class token* é prever a classe da imagem de entrada. Ambos são aprendidos pela rede por *back-propagation*. Ao final, a saída do modelo é obtida pelo resultado da função *softmax* da soma dos vetores *class token* e *distillation token*.

Touvron et al. (2021) realizaram experimentos com modelos CNN e *Transformer* para a utilização como professor, obtendo seus melhores resultados com o modelo RegNetY-16GF (RADOSAVOVIC et al., 2020), destacando que o fato deste modelo ser um professor melhor provavelmente se deve ao viés indutivo de CNN herdado por meio da destilação, e por generalizar melhor que modelos *Transformer* com as técnicas de aumento de dados utilizadas.

Figura 6 – Procedimento de destilação do modelo DeiT.



Fonte: (TOUVRON et al., 2021).

Assim como o ViT, o modelo DeiT também possui três variações, onde o que muda basicamente é a quantidade de cabeças no mecanismos de atenção multi-cabeça, conforme detalhado na Tabela 4. Entretanto, devido a seu desempenho superior foram projetados modelos menores comparados ao ViT, contudo obtendo resultados semelhantes.

Tabela 4 – Detalhamento das variações do modelo DeiT.

Modelo	**Blocos**	**Cabeças**	**Parâmetros**
DeiT-Tiny (DeiT-Ti)	12	3	5 milhões
DeiT-Small (DeiT-S)	12	6	22 milhões
DeiT-Base (DeiT-B)	12	12	86 milhões

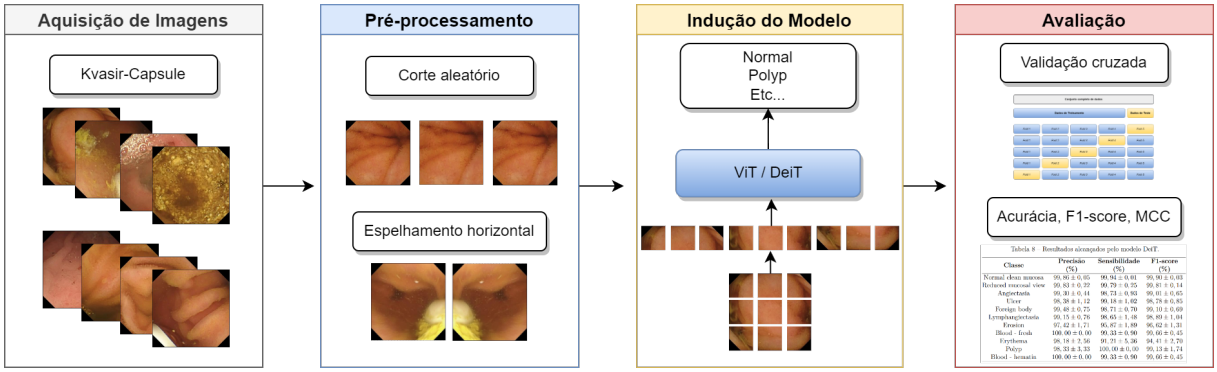
3.3 Considerações Finais

Neste capítulo foram apresentados os principais conceitos em que se baseia o método proposto neste trabalho, discorrendo sobre exames de cápsulas com a utilização de imagens geradas por WCEs e as tecnologias baseadas em *Transformer* empregadas na tarefa de reconhecimento de padrões em imagens. As definições apresentadas formam a base teórica para o entendimento da metodologia proposta no Capítulo 4.

4 Metodologia

Este capítulo apresenta a metodologia proposta para a classificação de imagens de exames de cápsulas endoscópicas, organizada segundo as seguintes etapas: aquisição de imagens provenientes de WCEs, pré-processamento, indução do modelo, e por fim sua avaliação. Os passos da metologia proposta são apresentados na Figura 7.

Figura 7 – Fluxograma da metodologia proposta.



Fonte: elaborado pelo autor.

4.1 Aquisição de Imagens

Para aquisição de imagens foi utilizado o *dataset* público Kvasir-Capsule (SMEDSRUD et al., 2021), composto por 117 vídeos obtidos por WCEs, coletados de exames realizados em hospitais da Noruega, totalizando mais de 4 milhões de *frames* extraídos, dos quais, após verificação criteriosa por quatro médicos especialistas, 47.238 *frames* foram rotulados e verificados, sendo classificados em 14 classes diferentes.

O Kvasir-Capsule é agrupado em duas categorias de imagens: *anatomy findings* e *luminal findings*. Em *anatomy findings* encontram-se as imagens relacionadas aos marcos anatômicos do trato gastrointestinal, tais como piloro, válvula ileocecal e ampola de Vater. Estes marcos são utilizados para orientação do especialista durante a realização de exame endoscópico. Já na categoria *luminal findings* foram agrupadas imagens relacionadas ao lúmen intestinal, podendo conter variações na mucosa intestinal, sangramentos, algum corpo estranho ou visibilidade reduzida (SMEDSRUD et al., 2021). Exemplos de todas as classes presentes neste *dataset* são apresentados na Figura 8.

Para atender ao objetivo proposto neste trabalho, foram utilizados somente exemplos da categoria *luminal findings*, por se tratarem de imagens que podem conter alguma anormalidade na mucosa. As classes que pertencem a esta categoria estão elencadas na listagem a seguir, bem como suas principais características:

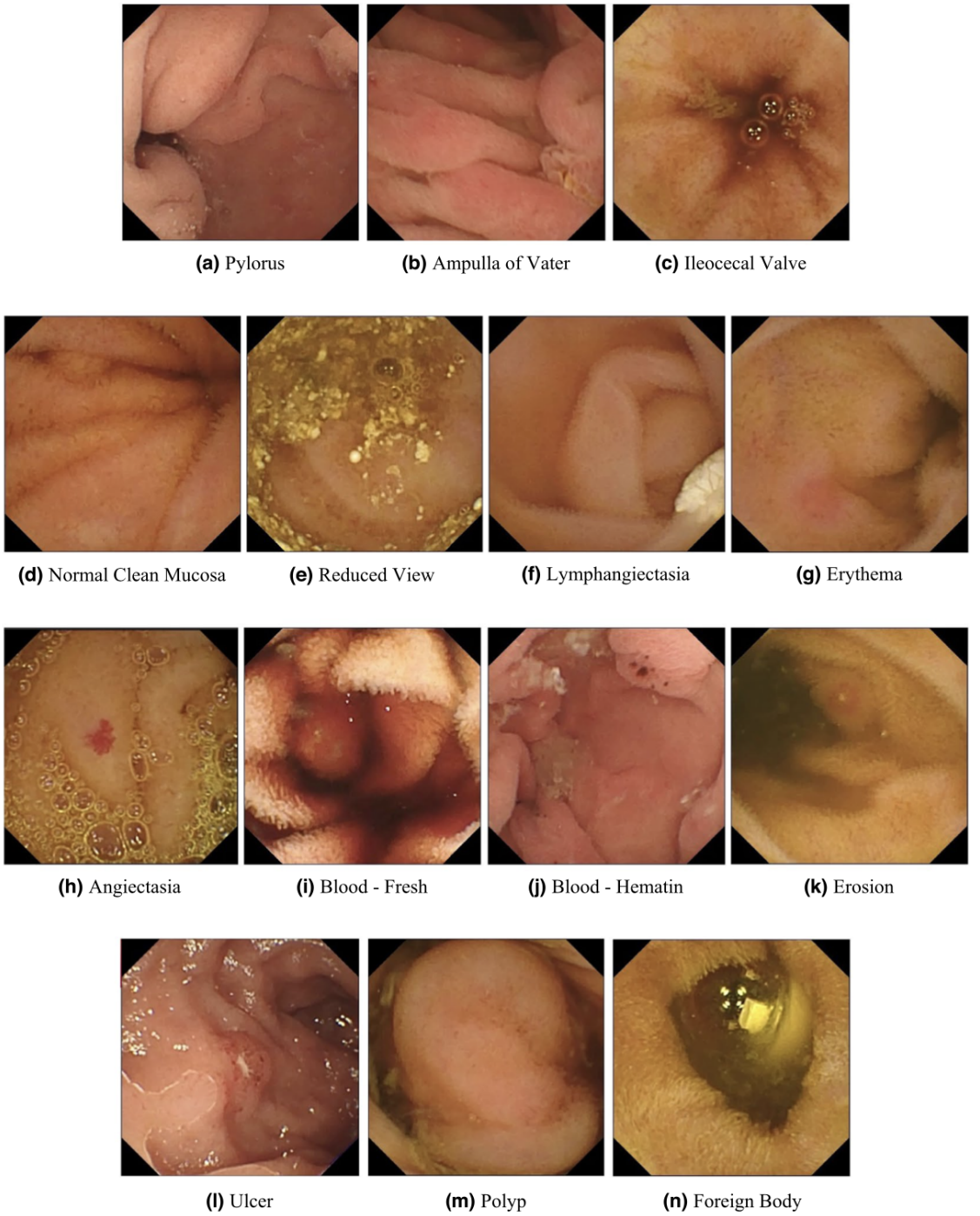
- *Normal clean mucosa*: exibe intestino delgado limpo com pouca ou nenhuma quantidade de fluido, mucosa com vilosidades saudáveis e sem achados patológicos;
- *Reduced mucosal view*: apresenta bolhas ou resíduos de fezes, prejudicando a visão da mucosa do intestino delgado;
- *Blood - fresh*: contém sangramento e líquido com coloração vermelha;
- *Blood - hematin*: exibe sangramento mínimo e pequenos pontos em tom vermelho escuro;
- *Foreign body*: apresenta algum corpo estranho na imagem, como resíduos de comprimidos ou cápsulas;
- *Erythema*: contém alterações típicas da mucosa com uma aparência avermelhada, chamada de mucosa eritematosa;
- *Angiectasias*: exibe pequenos vasos superficiais dilatados com sangramento crônico, que posteriormente podem causar anemia;
- *Erosion*: apresenta pequenas lesões escavadas que erodem em diferentes extensões na superfície da mucosa;
- *Ulcer*: exibe grandes lesões na mucosa que podem causar a estenose do lúmen, dificultando a absorção e passagem de nutrientes;
- *Lymphangiectasia*: exibe vasos linfóides dilatados na parede da mucosa;
- *Polyp*: apresenta saliências da parede mucosa contendo lesões pré-cancerosas.

Conforme pode-se observar na Tabela 5, este conjunto de imagens é altamente desbalanceado. Este é um desafio global na área médica porque alguns achados são mais comuns que outros, o que adiciona um desafio para os pesquisadores, uma vez que os métodos aplicados aos dados também devem ser capazes de aprender com uma quantidade pequena de dados de treinamento (SMEDSRUD et al., 2021).

4.2 Pré-processamento

Os modelos *Transformer* utilizados neste trabalho operam com um tamanho fixo da imagem de entrada, dessa forma a etapa de pré-processamento inicia-se com o redimensionamento das imagens para resolução esperada pelos modelos, a saber 224x224 *pixels*. Destaca-se que para o conjunto de treinamento é aplicado o corte e redimensionamento aleatório das imagens. Seu funcionamento se dá seguinte forma: primeiro é calcula a área de corte da imagem, variando de 75% até 100% da área original, em seguida a área recortada

Figura 8 – Exemplos das 14 classes do *dataset* Kvasir-Capsule. Imagens de **(a)** a **(c)** correspondem à categoria *anatomy findings*, enquanto exemplos de **(d)** a **(n)** pertencem à categoria *luminal findings*.



Fonte: (SMEDSRUD et al., 2021).

é redimensionada para a resolução esperada, ou seja, primeiro é aplicado um corte de área e posição aleatória na imagem, sendo em seguida a área recortada ajustada para a resolução de 224x224 *pixels*. No conjunto de teste não há corte aleatório, apenas o ajuste para o tamanho esperado aplicado na imagem original.

Tabela 5 – Distribuição da quantidade de exemplos por classe e da proporção ao total de imagens do *dataset* Kvasir-Capsule.

Classe	**Quantidade amostras**	**Proporção**
Normal clean mucosa	34338	82,72%
Reduced mucosal view	2906	7,00%
Angiectasia	866	2,09%
Ulcer	854	2,06%
Foreign body	776	1,87%
Lymphangiectasia	592	1,43%
Erosion	507	1,22%
Blood - fresh	446	1,07%
Erythema	159	0,38%
Polyp	55	0,13%
Blood - hematin	12	0,03%

No conjunto de treinamento é aplicado também o espelhamento horizontal. Nesta técnica as imagens são espelhadas no eixo horizontal, sendo utilizada proporção de 50% para aplicação nas imagens, ou seja, do montante total de imagens, na metade dos exemplos foi aplicado espelhamento horizontal. Convém frisar que esta técnica não foi utilizada no conjunto de teste, apenas no conjunto de treinamento.

Por fim, todas as imagens passam por normalização dos canais RGB, com média e desvio padrão de acordo com os parâmetros ideais para cada modelo, especificados conforme seus autores. Para o modelo ViT, são utilizados os valores de [0.5, 0.5, 0.5] tanto para média quanto para desvio padrão. Já para o modelo DeiT são utilizados os valores padrão do *dataset* ImageNet, que são [0.485, 0.456, 0.406] para média e [0.229, 0.224, 0.225] para desvio padrão. Estes valores foram calculados conforme o *dataset* utilizado por cada modelo em seu pré-treinamento. A normalização se dá da seguinte forma: os valores dos *pixels* de todos os canais RGB são transformados do intervalo de números inteiros de 0 a 255 para o intervalo de números decimais de 0 a 1, em seguida estes valores são subtraídos da média e divididos pelo desvio padrão. Este procedimento pode ser sintetizado na Equação 4.1.

$$Normalizacao(R,G,B) = \frac{Entrada(R,G,B) - Media(R,G,B)}{DesvioPadrao(R,G,B)} \tag{4.1}$$

4.3 Indução do Modelo

Conforme detalhado no Capítulo 3, para se atingir bons resultados com modelos *Transformer*, estes devem ser treinados em grandes bases de dados. Portando, a sistemática para utilização destes modelos é que se faça um pré-treinamento em um *dataset* grande, e depois um ajuste fino no *dataset* alvo da classificação. Neste trabalho foram realizados

experimentos com as arquiteturas ViT e DeiT, visando a classificação das 11 classes da categoria *luminal findings* do *dataset* Kvasir-Capsule.

Conforme visto anteriormente, o modelo DeiT é uma evolução do modelo ViT, conseguindo atingir resultados muito significativos ao utilizar um *dataset* bem menor para treinamento. O modelo DeiT utiliza em seu treinamento abordagem com destilação de conhecimento, onde é utilizada a rede CNN RegNetY-16GF como professor. Esta abordagem permite ao *Transformer* descobrir com eficiência representações úteis para as imagens de entrada.

Estes modelos utilizam conduta semelhante à utilizada em tarefas de processamento de linguagem natural, onde uma sequência de palavras é quebrada em vários *tokens*, sendo em seguida gerado o *embedding* destes *tokens* para ser alimentado o codificador *Transformer*. Para a classificação das imagens, inicialmente divide-se a imagem a ser analisada em 196 *patches* de tamanho 16x16 *pixels*, em seguida é gerada a sequência de *patch tokens*, para juntamente com o *positional encoding*, o *class token*, e o *distillation token* no caso específico do modelo DeiT, formarem a projeção linear utilizada como entrada para o codificador *Transformer*. Portanto, os *patches* da imagem de entrada são tratados da mesma forma que os *tokens* são utilizados em tarefas relacionadas a NLP.

Para auxiliar na etapa de ajuste fino foi utilizada a biblioteca *Transformers*[1]. Esta biblioteca facilita a utilização de modelos pré-treinados publicados na plataforma *Hugging Face*[2], provendo uma interface simplificada para inicializar, treinar e avaliar tais modelos.

Após o ajuste fino realizado no *dataset* Kvasir-Capsule, é efetuada a classificação na camada MLP final, chamada MLP *head*, composta por camada linear de dimensão $D \times K$, onde D é a dimensão dos vetores do *Transformer* e K a quantidade de classes, sendo em seguida aplicada função *softmax*.

4.4 Avaliação

Para avaliação de resultados foram utilizadas as métricas acurácia, precisão, sensibilidade e *f1-score* como critérios para avaliação do modelo proposto. Para explicar os indicadores utilizados neste trabalho, utiliza-se a matriz de confusão observada na Tabela 6. Esta tabela é um resumo dos valores obtidos pelo algoritmo de classificação, dando uma visão geral do que o modelo acertou e quais erros cometeu. De uma forma geral, os quatro valores possíveis da matriz de confusão podem ser sintetizados da seguinte forma:

- Verdadeiro Positivo (VP): referem-se aos exemplos que pertencem à classe analisada e que o modelo previu corretamente como pertencente a esta classe;

[1] <https://huggingface.co/docs/transformers/index>
[2] <https://huggingface.co/>

- Verdadeiro Negativo (VN): referem-se aos exemplos que não pertencem à classe analisada e o modelo previu corretamente como não pertencente a esta classe;
- Falso Positivo (FP): referem-se aos exemplos que não pertencem à classe analisada, mas o modelo previu erroneamente como pertencente a esta classe;
- Falso Negativo (FN): referem-se aos exemplos que pertencem à classe analisada, mas o modelo previu erroneamente como não pertencente a esta classe.

Tabela 6 – Matriz de confusão.

		Valor verdadeiro	
		Positivo	Negativo
Valor predito	Positivo	VP	FP
	Negativo	FN	VN

Segundo Marsland (2014), a acurácia (Equação 4.2) representa a proporção de acertos do modelo, independente da classe. Pega-se o número total de observações que o modelo acertou e divide-se pelo número total de observações que o modelo previu. Informa quantas amostras foram classificadas corretamente.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.2}$$

A sensibilidade (Equação 4.3), também chamada revocação, ou *recall*, aponta a proporção de casos positivos identificados corretamente. Para calculá-la, toma-se o número de observações que o modelo classificou como positivos corretamente, os verdadeiros positivos (VP) e divide-se pelo número total de observações com rótulo positivo, verdadeiros positivos (VP) e falsos negativos (FN) (MARSLAND, 2014).

$$Sensibilidade = \frac{VP}{VP + FN} \tag{4.3}$$

A precisão (Equação 4.4) é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos, ou seja, dos exemplos classificados como positivos, indica quantos realmente são positivos (MARSLAND, 2014).

$$Precisão = \frac{VP}{VP + FP} \tag{4.4}$$

O *f1-score* (Equação 4.5) é a média ponderada da precisão e sensibilidade. Leva em conta tanto os falsos positivos (FP) quanto os falsos negativos (FN). Quanto maior o valor, mais precisa foi a classificação realizada (MARSLAND, 2014).

$$F1\text{-}score = 2 * \frac{Precisão * Sensibilidade}{Precisão + Sensibilidade} \tag{4.5}$$

Além disso, foi utilizada também a métrica MCC (Equação 4.6), do inglês *Matthews Correlation Coefficient*, que segundo Chicco e Jurman (2020), é uma métrica mais apropriada ao se avaliar *datasets* desbalanceados do que *f1-score*, pois leva em consideração todos os quatro valores da matriz de confusão, e um valor alto significa que ambas as classes são bem previstas, mesmo que uma classe esteja desproporcionalmente representada.

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{4.6}$$

Para avaliar a generalização do modelo foi realizada validação cruzada *K-Fold*. Nesta técnica a base de dados de amostras é subdivida em k subconjuntos, para cada k rodada de execução de treinamento e teste do modelo. A cada rodada, um subconjunto diferente é usado como teste, até que todos os subconjuntos tenham sido testados. Os subconjuntos restantes são combinados para formar um conjunto de treinamento. O resultado final é encontrado calculando-se a média dos resultados obtidos na validação cruzada, assim como o desvio padrão. A Figura 9 apresenta exemplo da divisão do conjunto de dados com validação cruzada em 5 *folds*.

Figura 9 – Exemplo de divisão do conjunto de dados com validação cruzada em 5 *folds*.

Fonte: elaborado pelo autor.

4.5 Considerações Finais

Neste capítulo foram apresentados as etapas da metodologia proposta, detalhando a composição do *dataset* utilizado, explicando a etapa de pré-processamento das imagens, assim como os procedimentos utilizados na indução e avaliação do modelo. No Capítulo 5 serão detalhados os experimentos realizados e valores obtidos, bem como a discussão dos resultados e comparação com outros trabalhos.

5 Resultados

Neste capítulo serão apresentados os resultados alcançados após a execução da metodologia proposta. Conforme tratado no Capítulo 4, foi realizada a classificação das 11 classes da categoria *luminal findings* do *dataset* Kvasir-Capsule, utilizando os modelos ViT e DeiT. Os experimentos foram executados na plataforma Kaggle[1], sendo utilizado ambiente com GPU para se obter melhor desempenho no treinamento e inferência de resultados. No conjunto de exemplos de treino, foram utilizadas técnicas de aumento de dados, sendo aplicados redimensionamento com corte e espelhamento horizontal de forma aleatória nas imagens.

Visando mensurar a generalização do método proposto, foi realizada validação cruzada estratificada em 5 *folds*. Dessa forma, para cada modelo avaliado, o treinamento foi executado cinco vezes, em cada execução com um conjunto de imagens diferentes, perfazendo uma distribuição de 80% para treino e 20% para teste. Por se tratar de validação cruzada estratificada, o *dataset* foi dividido em cinco blocos com proporções semelhantes para todas as classes analisadas.

Para avaliação dos modelos foram utilizadas as métricas acurácia, precisão, sensibilidade e *f1-score*. Os valores apresentados nas tabelas de resultados foram obtidos calculando-se a média das 5 execuções da validação cruzada, bem como do desvio padrão, demonstrando a variação esperada no desempenho dos modelos.

5.1 Experimentos com Modelos Transformer

A primeira etapa dos experimentos se deu com o ajuste fino do modelo ViT. Seu treinamento foi realizado por 4 épocas, sendo a avaliação realizada a cada 100 passos. A quantidade de passos por época é calculada dividindo-se o total de exemplos de treinamento pelo valor do *batch size*. A variação de modelo escolhida foi o ViT-B[2], pré-treinado no *dataset* ImageNet-21k, composto por 14 milhões de imagens agrupadas em 21 mil classes diferentes. Os parâmetros de configuração utilizados foram os seguintes: 5 *folds*, 4 épocas, *batch size* de 32, *learning rate* de 0,0002 e otimizador *AdamW*. Na Tabela 7 estão listados valores médios e desvio padrão obtidos nas métricas precisão, sensibilidade ou *recall* e *f1-score* para todas as 11 classes avaliadas.

Em seguida foi executado o experimento com o modelo DeiT, utilizando os mesmos parâmetros de configuração, técnicas de aumento de dados e divisão do *dataset* em 5 *folds*

[1] <https://www.kaggle.com/>
[2] <https://huggingface.co/google/vit-base-patch16-224-in21k>

Tabela 7 – Resultados alcançados pelo modelo ViT.

Classe	**Precisão** (%)	**Sensibilidade** (%)	**F1-score** (%)
Normal clean mucosa	$99,83 \pm 0,06$	$99,93 \pm 0,04$	$99,88 \pm 0,02$
Reduced mucosal view	$99,79 \pm 0,20$	$99,90 \pm 0,14$	$99,85 \pm 0,06$
Angiectasia	$99,19 \pm 0,86$	$98,04 \pm 1,13$	$98,60 \pm 0,75$
Ulcer	$97,94 \pm 1,82$	$98,48 \pm 1,36$	$98,20 \pm 1,23$
Foreign body	$99,48 \pm 0,63$	$98,45 \pm 0,77$	$98,96 \pm 0,52$
Lymphangiectasia	$99,49 \pm 0,68$	$98,48 \pm 1,64$	$98,98 \pm 0,92$
Erosion	$94,67 \pm 2,69$	$92,91 \pm 4,31$	$93,70 \pm 2,36$
Blood - fresh	$99,56 \pm 0,89$	$99,55 \pm 0,55$	$99,55 \pm 0,65$
Erythema	$96,20 \pm 3,11$	$91,81 \pm 3,21$	$93,89 \pm 1,87$
Polyp	$100,00 \pm 0,00$	$98,18 \pm 3,64$	$99,05 \pm 1,90$
Blood - hematin	$90,00 \pm 20,00$	$100,00 \pm 0,00$	$93,33 \pm 13,13$

utilizados anteriormente. A variante de modelo empregada foi o DeiT-B[3], pré-treinado no *dataset* ImageNet-1k, composto por 1 milhão de imagens agrupadas em mil classes diferentes. A Tabela 8 apresenta os valores médios e desvio padrão alcançados pelo modelo DeiT em todas as classes avaliadas.

Tabela 8 – Resultados alcançados pelo modelo DeiT.

Classe	**Precisão** (%)	**Sensibilidade** (%)	**F1-score** (%)
Normal clean mucosa	$99,86 \pm 0,05$	$99,94 \pm 0,01$	$99,90 \pm 0,03$
Reduced mucosal view	$99,83 \pm 0,22$	$99,79 \pm 0,25$	$99,81 \pm 0,14$
Angiectasia	$99,30 \pm 0,44$	$98,73 \pm 0,93$	$99,01 \pm 0,65$
Ulcer	$98,38 \pm 1,12$	$99,18 \pm 1,02$	$98,78 \pm 0,85$
Foreign body	$99,48 \pm 0,75$	$98,71 \pm 0,70$	$99,10 \pm 0,69$
Lymphangiectasia	$99,15 \pm 0,76$	$98,65 \pm 1,48$	$98,89 \pm 1,04$
Erosion	$97,42 \pm 1,71$	$95,87 \pm 1,89$	$96,62 \pm 1,31$
Blood - fresh	$100,00 \pm 0,00$	$99,33 \pm 0,90$	$99,66 \pm 0,45$
Erythema	$98,18 \pm 2,56$	$91,21 \pm 5,36$	$94,41 \pm 2,70$
Polyp	$98,33 \pm 3,33$	$100,00 \pm 0,00$	$99,13 \pm 1,74$
Blood - hematin	$100,00 \pm 0,00$	$99,33 \pm 0,90$	$99,66 \pm 0,45$

Comparando os valores obtidos por ambos os modelos, observamos que o modelo DeiT obteve resultados ligeiramente melhores em *f1-score* na maioria das classes. Apenas nas classes *reduced mucosal view* e *lymphangiectasia* o modelo ViT se saiu levemente melhor que o DeiT. Um ponto de destaque foi o resultado alcançado pelos modelos ao analisar imagens da classe *blood - hematin*, com apenas 12 exemplos presentes no *dataset*, onde o modelo ViT registrou um desvio padrão muito alto, ao contrário do modelo DeiT, evidenciando a melhora nos resultados com o uso da técnica de destilação de conhecimento, fazendo com que o modelo *Transformer* herdasse vieses indutivos do modelo professor.

[3] <https://huggingface.co/facebook/deit-base-distilled-patch16-224>

Ao final da execução dos 5 *folds*, o modelo ViT obteve taxas médias de 99,68% para acurácia, 97,83% para precisão, 97,79% para sensibilidade e 97,64% para *f1-score*, enquanto o modelo DeiT alcançou médias de 99,75% para acurácia, 98,17% para precisão, 98,31% para sensibilidade e 98,06% para *f1-score*.

5.2 Experimentos com Modelos CNN

Outros experimentos foram realizados com o objetivo de comprovar a consistência do método proposto comparando resultados com arquiteturas CNN bastante utilizadas na classificação de imagens obtidas por WCE, conforme destaca o trabalho de Muruganantham e Balakrishnan (2021). Os modelos escolhidos foram o ResNet-50 (HE et al., 2015) e o DenseNet-121 (HUANG et al., 2017) pré-treinados no *dataset* ImageNet.

Foi utilizada a mesma sistemática dos experimentos anteriores realizados com os modelos *Transformer*, ou seja, foi utilizada a mesma base dados com validação cruzada em 5 *folds* e técnicas de aumento de dados no conjunto de treino, sendo aplicados redimensionamento com corte e espelhamento horizontal de forma aleatória nas imagens. A avaliação de tais modelos se deu ao final de cada época. Os parâmetros utilizados no treinamento foram os seguintes: 5 *folds*, 50 épocas, *batch size* de 32, otimizador *Adam* e *learning rate* partindo de 0,001 até 0,0002, fazendo uso da função de retorno *ReduceLROnPlateau* que reduz o *learning rate* sempre que o aprendizado fica estagnado. Nas Tabelas 9 e 10 estão listados valores médios e desvio padrão para todas as classes analisadas.

Tabela 9 – Resultados obtidos pelo modelo ResNet-50.

Classe	**Precisão** (%)	**Sensibilidade** (%)	**F1-score** (%)
Normal clean mucosa	$99,42 \pm 0,05$	$99,86 \pm 0,08$	$99,64 \pm 0,04$
Reduced mucosal view	$99,28 \pm 0,42$	$99,31 \pm 0,31$	$99,29 \pm 0,30$
Angiectasia	$98,82 \pm 1,12$	$95,61 \pm 1,07$	$97,18 \pm 0,71$
Ulcer	$98,25 \pm 1,01$	$97,66 \pm 2,03$	$97,94 \pm 1,09$
Foreign body	$98,56 \pm 0,95$	$95,49 \pm 1,82$	$96,98 \pm 0,67$
Lymphangiectasia	$99,12 \pm 0,95$	$94,26 \pm 1,72$	$96,62 \pm 1,15$
Erosion	$96,36 \pm 3,14$	$85,40 \pm 3,50$	$90,48 \pm 2,13$
Blood - fresh	$99,56 \pm 0,89$	$97,53 \pm 1,31$	$98,52 \pm 0,59$
Erythema	$89,00 \pm 11,53$	$86,21 \pm 9,36$	$86,76 \pm 7,29$
Polyp	$100,00 \pm 0,00$	$100,00 \pm 0,00$	$100,00 \pm 0,00$
Blood - hematin	$100,00 \pm 0,00$	$100,00 \pm 0,00$	$100,00 \pm 0,00$

Importante destacar que a quantidade de épocas para treinamento dos modelos foi calculada de forma a não atingir o limite máximo de 10 horas do tempo disponibilizado pela plataforma Kaggle para a execução dos experimentos. Dessa forma, o tempo de execução dos experimentos dos modelos DeiT e ViT quase atingiu o limite de 10 horas, enquanto os

Tabela 10 – Resultados obtidos pelo modelo DenseNet-121.

Classe	**Precisão** (%)	**Sensibilidade** (%)	**F1-score** (%)
Normal clean mucosa	$99,14 \pm 0,27$	$99,76 \pm 0,02$	$99,45 \pm 0,14$
Reduced mucosal view	$99,38 \pm 0,58$	$98,83 \pm 0,55$	$99,10 \pm 0,24$
Angiectasia	$97,99 \pm 1,06$	$95,04 \pm 0,69$	$96,48 \pm 0,57$
Ulcer	$96,85 \pm 1,27$	$96,61 \pm 1,19$	$96,72 \pm 0,93$
Foreign body	$97,33 \pm 1,10$	$93,69 \pm 2,01$	$95,46 \pm 1,34$
Lymphangiectasia	$96,82 \pm 1,41$	$92,41 \pm 3,10$	$94,54 \pm 1,98$
Erosion	$94,91 \pm 3,14$	$78,48 \pm 9,10$	$85,46 \pm 5,11$
Blood - fresh	$100,00 \pm 0,00$	$97,98 \pm 1,65$	$98,97 \pm 0,85$
Erythema	$92,63 \pm 5,13$	$79,33 \pm 9,54$	$84,91 \pm 4,16$
Polyp	$98,33 \pm 3,33$	$98,18 \pm 3,64$	$98,18 \pm 2,24$
Blood - hematin	$100,00 \pm 0,00$	$100,00 \pm 0,00$	$100,00 \pm 0,00$

experimentos dos modelos ResNet-50 e DenseNet-121 foram executados durante pouco mais de 9 horas e meia.

Ao final dos experimentos o modelo ResNet-50 obteve taxas médias de 99,27% para acurácia, 98,03% para precisão, 95,58% para sensibilidade e 96,67% para *f1-score*. Já o modelo DenseNet-121 alcançou médias de 98,95% para acurácia, 97,58% para precisão, 93,66% para sensibilidade e 95,39% para *f1-score*.

5.3 Discussão

Ao se analisar os dados contidos nas Tabelas 7, 8, 9 e 10, pode-se observar resultados similares, entretanto, com dados mais robustos obtidos pelos modelos *Transformer*, e em especial com o modelo DeiT, alcançando altas taxas de acurácia, precisão, sensibilidade e *f1-score* em todas as classes analisadas, sempre associadas a um baixo desvio padrão. Evidencia-se os valores obtidos pelo modelo ResNet nas classes *polyp* e *blood - hematin*, acertando todos os exemplos testados. Estas classes são as que apresentam o menor número de amostras, apenas 55 e 12, respectivamente, conforme pode ser observado na Tabela 5.

A Tabela 11 apresenta comparação dos valores médios e desvio padrão obtidos pelos modelos DeiT, ViT, ResNet-50 e DenseNet-121. Ressalta-se que o modelo DeiT superou os demais modelos em todos os critérios de avaliação, apresentando resultados levemente superiores ao modelo ViT. Já o modelo DenseNet-121 não figurou com melhor resultado em nenhuma métrica avaliada, apesar de alcançar valores próximos aos demais modelos. Além disso, foram comparados também os valores para MCC pois, como o *dataset* utilizado é altamente desbalanceado, um valor alto para esta métrica significa que todas as classes são bem previstas, mesmo que uma classe esteja desproporcionalmente representada. Os valores obtidos em MCC evidenciam novamente a superioridade do modelo DeiT em comparação aos demais.

Tabela 11 – Comparativo de valores obtidos pelos modelos DeiT, ViT, ResNet-50 e DenseNet-121.

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)	F1-score (%)	MCC (%)
DeiT	$\mathbf{99,75 \pm 0,07}$	$\mathbf{98,17 \pm 6,77}$	$\mathbf{98,31 \pm 3,12}$	$\mathbf{98,06 \pm 4,71}$	$\mathbf{99,20 \pm 0,23}$
ViT	$99,68 \pm 0,05$	$97,83 \pm 6,87$	$97,79 \pm 3,40$	$97,64 \pm 4,89$	$98,98 \pm 0,16$
ResNet-50	$99,27 \pm 0,11$	$98,03 \pm 4,75$	$95,58 \pm 5,92$	$96,67 \pm 4,69$	$97,63 \pm 0,36$
DenseNet-121	$98,95 \pm 0,24$	$97,58 \pm 3,08$	$93,66 \pm 8,51$	$95,39 \pm 5,56$	$96,58 \pm 0,79$

Embora o modelo DeiT apresente resultados superiores, os valores obtidos pelo modelo ViT são semelhantes. Isto posto, visando avalizar a utilização de um modelo em relação ao outro, foi realizada análise de significância estatística (COX, 1982) entre os resultados obtidos por estes modelos. A hipótese nula definida ocorre quando não há diferença significativa entre as métricas observadas. O nível de significância estabelecido para o teste é $\alpha = 0{,}05$. Se o valor de p for menor que α, a hipótese nula é descartada, o que indica que há diferença significativa entre os resultados alcançados, confirmando dessa maneira a superioridade do modelo DeiT. A Tabela 12 apresenta o resultado da análise, demonstrando que as métricas precisão, sensibilidade, *f1-score* e MCC obtidas pelo modelo DeiT apresentam significância estatística frente às alcançadas pelo modelo ViT.

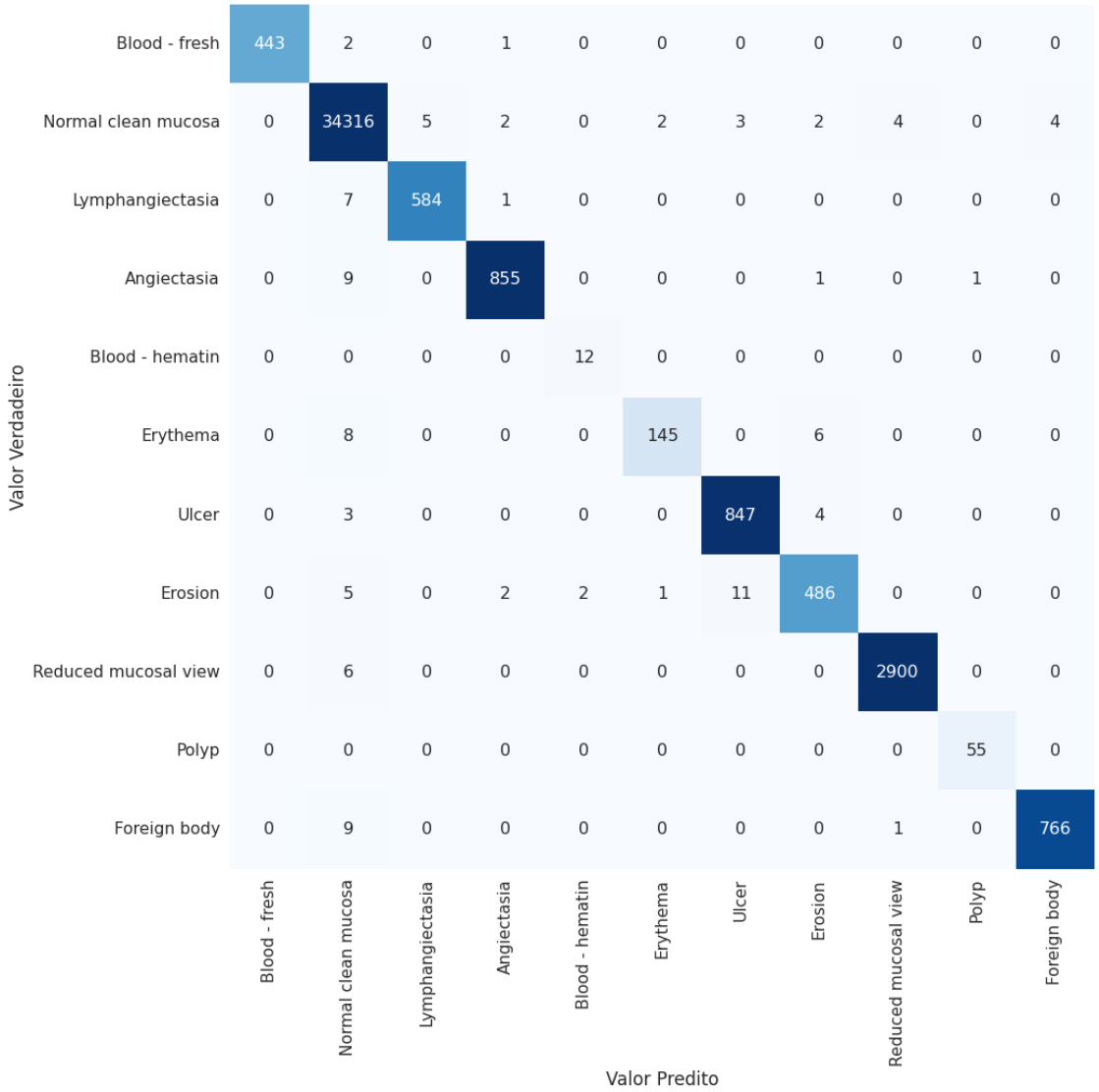
Tabela 12 – Análise de significância estatística nas métricas alcançadas pelos modelos ViT e DeiT.

	Acurácia	Precisão	Sensibilidade	F1-score	MCC
ViT	0,9968	0,9783	0,9779	0,9764	0,9898
DeiT	0,9975	0,9817	0,9831	0,9806	0,9920
Valor de p	0,0585	0,0005	0,0000	0,0000	0,0008
Resultado	Não significante	Significante	Significante	Significante	Significante

Apesar dos promissores resultados alcançados, o modelo DeiT apresentou algumas falhas. Ao final do experimento, após as 41511 imagens terem sido testadas através da execução de validação cruzada em 5 *folds*, ainda ocorreram 102 falhas (Figura 10), perfazendo uma taxa de 0,25% de erros do modelo.

A maioria das falhas, 49 no total, ocorreram ao modelo predizer erroneamente uma imagem como pertencente à classe *normal clean mucosa*. Uma possível causa deste problema se deve ao fato desta classe ser a que possui mais exemplos, conforme pode ser visto na Tabela 5, o que pode ter induzido o modelo a privilegiar mais esta classe em detrimento a outras. Outra hipótese para a ocorrências destas falhas é a semelhança entre as imagens analisadas. Como pode ser observado na Figura 11, os exemplos **(a)**, **(c)**, **(d)** são muito parecidos, sendo estes rotulados no *dataset* como pertencentes à classe *normal clean mucosa*, mas classificados pelo modelo como *foreign body*. Tomando como base a descrição das classes contidas no Capítulo 4, estes exemplos poderiam ser confundidos também como pertencentes à classe *reduced mucosal view* devido à presença de bolhas

Figura 10 – Matriz de confusão dos resultados do modelo DeiT.



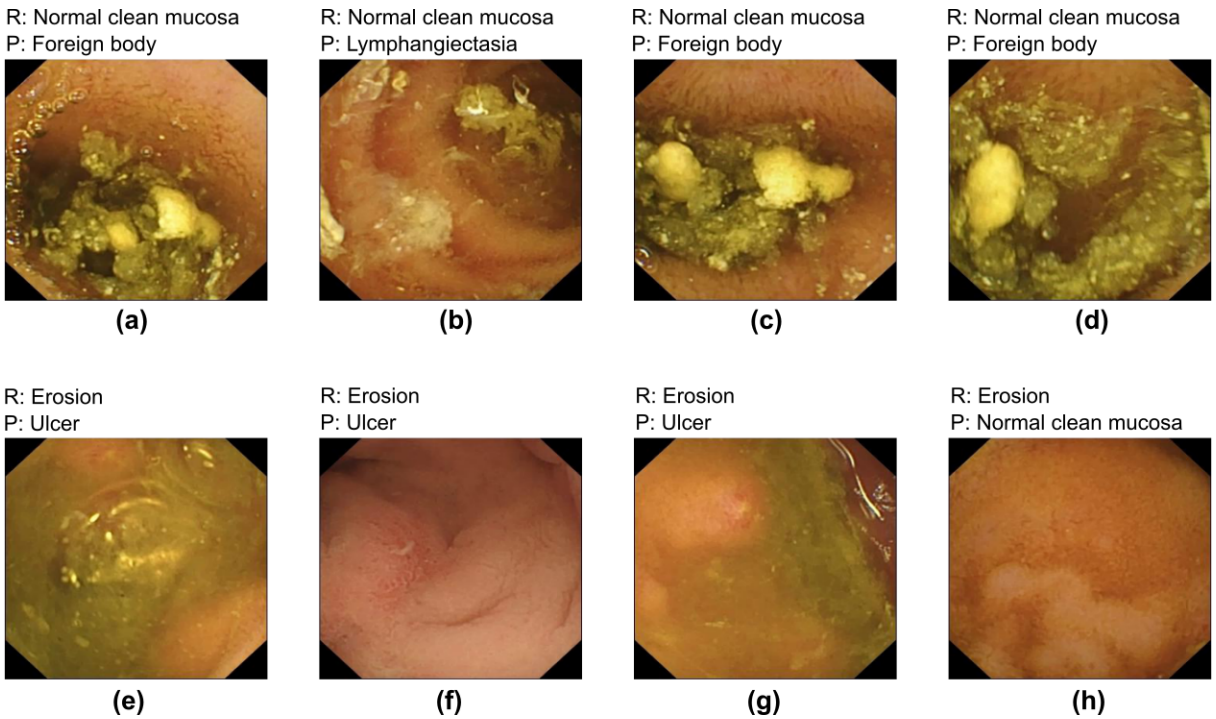
Fonte: elaborado pelo autor.

nas imagens. As imagens pertencentes à classe *erosion* abrangem outro exemplo onde o modelo DeiT apresentou algumas falhas, pois, devido a presença de pequenas lesões na mucosa, foram erroneamente classificadas como *ulcer* e *normal clean mucosa*, conforme os exemplos de **(e)** a **(h)**.

5.4 Comparação com Outros Trabalhos

A Tabela 13 apresenta comparativo de trabalhos que focaram na classificação de anomalias do trato gastrointestinal utilizando o *dataset* Kvasir-Capsule. Esta comparação se mostrou uma tarefa árdua, uma vez que as metodologias, conjunto de imagens e quantidade de amostras foram bastante diferentes entre os trabalhos analisados. Importante frisar que os resultados obtidos pelo método proposto nesta dissertação foram superiores aos demais

Figura 11 – Exemplo de falhas ocorridas no modelo DeiT. Legenda: R é o rótulo presente no *dataset*, P é a previsão do modelo.



Fonte: elaborado pelo autor.

trabalhos.

Muruganantham e Balakrishnan (2022) propuseram arquitetura híbrida intercalando blocos com camadas convolucionais e mecanismos de atenção para avaliar as imagens do *dataset* Kvasir-Capsule, entretanto utilizaram apenas 3 classes em seus experimentos, cada classe com 800 imagens, totalizando 2400 exemplos. Em seus experimentos foi utilizada validação cruzada em 4 *folds*, obtendo como melhores resultados 95,36% de acurácia, 95,20% para precisão e 95,25% para *f1-score*. Abordagem semelhante foi adotada por Srivastava et al. (2022) ao utilizarem arquitetura híbrida com camadas convolucionais e mecanismos de atenção, entretanto com resultado diverso. Foram utilizadas classes de ambas as categorias do *dataset*, descartando as 3 classes com menos exemplos, empregando no total 11 classes na classificação. Em seus experimentos, as imagens foram divididas em uma proporção incomum, sendo 49% para treino e 51% para teste. Ao final alcançaram resultado modesto, com 63,73% para acurácia, 75,57% para precisão, 63,73% para sensibilidade e 67,34% para *f1-score*.

No trabalho de Amiri, Hassanpour e Beghdadi (2021) foi empregada abordagem tradicional para extração de características de textura, cor e forma das imagens, utilizando ao final classificador SVM para obtenção de resultados. Salienta-se que foram utilizadas todas as classes da categoria *luminal findings*, com exceção da classe *blood - hematin* devido à pouca quantidade de exemplos. Em seus experimentos aplicaram validação cruzada em

10 *folds*, e para balanceamento do *dataset*, utilizaram apenas 1000 exemplos da classe *normal clean mucosa*. Ao final alcançaram médias de 89,60% para acurácia, 89,50% para precisão, 89,60% para sensibilidade e 89,50% para *f1-score*. Convém destacar que em todos estes trabalhos as arquiteturas foram treinadas do zero, ou seja, não utilizaram os pesos de um pré-treinamento realizado em outro *dataset*.

Gjestang et al. (2021) utilizaram tanto arquitetura CNN tradicional com o modelo EfficientNet quanto abordagem professor-aluno, utilizando modelos CNN tanto para o professor quanto para o aluno, sendo esta última a que obteve melhores resultados. Em seus experimentos foram utilizadas todas as 14 classes do *dataset* Kvasir-Capsule, em ambas as categorias. Entretanto, adotaram validação cruzada em apenas 2 *folds*, não obtendo resultados muito satisfatórios, com 69,50% de acurácia, 73,40% de precisão, 69,60% de sensibilidade e 70,40% de *f1-score*.

Por fim, no trabalho de Bai et al. (2022) foi abordado método semelhante ao adotado nesta dissertação, com a utilização do modelo ViT para classificação das 11 classes da categoria *luminal findings* do *dataset* Kvasir-Capsule. Entretanto, tal modelo foi treinado do zero. Conforme destacado no Capítulo 3, para atingir bons resultados o modelo ViT precisa ser treinado em um grande *dataset*, com milhões de imagens. Por conta disso, os resultados apresentados foram moderados, atingindo acurácia de 79,15%.

Tabela 13 – Comparativo de valores obtidos por diferentes métodos.

Modelo	**Acurácia** (%)	**Precisão** (%)	**Sensibilidade** (%)	**F1-score** (%)
Método proposto (DeiT)	**99,75**	**98,17**	**98,31**	**98,06**
Método proposto (ViT)	99,68	97,83	97,79	97,64
(MURUGANANTHAM; BALAKRISHNAN, 2022)	95,36	95,20	-	95,25
(AMIRI; HASSANPOUR; BEGHDADI, 2021)	89,60	89,50	89,60	89,50
(GJESTANG et al., 2021)	69,50	73,40	69,60	70,40
(SRIVASTAVA et al., 2022)	63,73	75,57	63,73	67,34
(BAI et al., 2022)	79,15	-	-	-

6 Conclusão

As doenças inflamatórias intestinais apresentam alta taxa de incidência na população, sendo umas das principais causas de internação hospitalar. O diagnóstico precoce de anomalias presentes na mucosa intestinal é de fundamental importância para eficácia do tratamento. O exame por meio de cápsulas endoscópicas sem fio se caracteriza por ser uma técnica não invasiva que visualiza todo o trato gastrointestinal do paciente sem causar desconforto. Entretanto, a investigação manual das imagens produzidas em tal exame se constitui um procedimento tedioso e propenso a erros, pois os vídeos geralmente podem atingir até 10 horas de duração, demandando alta concentração do especialista médico na realização do diagnóstico.

Técnicas de aprendizado de máquina têm sido aplicadas com sucesso no desenvolvimento de sistemas de diagnóstico auxiliados por computador, onde, na última década, as Redes Neurais Convolucionais (CNNs) tornaram-se muito bem-sucedidas no reconhecimento de padrões em imagens. Atualmente, arquiteturas baseadas em *Transformer* constituem o estado da arte nas tarefas relacionadas a processamento de linguagem natural. Motivado por este sucesso, surgiram novas arquiteturas utilizando tal tecnologia visando resolver tarefas pertinentes a visão computacional.

Neste trabalho foi apresentada uma abordagem para a identificação de anormalidades presentes em imagens extraídas de vídeos de exames de endoscopia por cápsula, utilizando arquitetura baseada em *Transformer*, visando auxiliar o especialista médico a atingir maior eficiência no diagnóstico. Foram utilizados os modelos ViT e DeiT, onde este último emprega procedimento de destilação do conhecimento na etapa de treinamento, visando agregar informações extraídas tanto pelo professor, baseado em arquitetura CNN, quanto pelo aluno, que utiliza arquitetura *Transformer*.

Os resultados obtidos comprovam a eficácia da abordagem desenvolvida, evidenciando a superioridade do modelo DeiT frente tanto ao modelo ViT quanto a arquiteturas CNN frequentemente utilizadas na análise de imagens obtidas por WCE, como a ResNet e a DenseNet. Ao final dos experimentos, onde foi realizada validação cruzada em 5 *folds*, utilizando o *dataset* Kvasir-Capsule, o modelo DeiT alcançou valores médios de 99,75% para acurácia, 98,17% para precisão, 98,31% para sensibilidade, 98,06% para f1-score e 99,20% para MCC.

Como trabalhos futuros, pretende-se avaliar a metodologia proposta em outros *datasets* de imagens WCE, assim como estressar o modelo em um conjunto de *datasets* contendo diversidade populacional na obtenção dos exames. Propõe-se ainda testar a coerência temporal, analisando os *frames* imediatamente anteriores e posteriores às falhas

encontradas, visando levantar hipóteses para o ocorrido. Por fim, intenciona-se aplicar modelos *Transformer* em tarefas de sumarização de vídeo, visando prover ao final soluções que facilitem o diagnóstico de doenças inflamatórias intestinais.

Esta dissertação gerou uma publicação de seus resultados científicos, conforme informado na Tabela 14. O método proposto foi publicado no 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), evento Qualis A1 em Ciência da Computação.

Tabela 14 – Artigo publicado em evento científico relacionado ao tema do diagnóstico de anomalias do trato gastrointestinal em imagens de endoscopia por cápsula.

Título	**Congresso/Conferência**	**Qualis**
Classification of Video Capsule Endoscopy Images Using Visual Transformers (LIMA et al., 2022)	IEEE-EMBS BHI 2022	A1

Referências

ALI, H.; SHARIF, M.; YASMIN, M.; REHMANI, M. H.; RIAZ, F. A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract. *Artificial Intelligence Review*, Springer, v. 53, n. 4, p. 2635–2707, 2020. Citado na página 15.

AMIRI, Z.; HASSANPOUR, H.; BEGHDADI, A. A computer-aided method for digestive system abnormality detection in wce images. *Journal of Healthcare Engineering*, Hindawi, v. 2021, 2021. Citado 2 vezes nas páginas 49 e 50.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. Citado na página 27.

BAI, L.; WANG, L.; CHEN, T.; ZHAO, Y.; REN, H. Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset. *Electronics*, MDPI, v. 11, n. 17, p. 2747, 2022. Citado 3 vezes nas páginas 21, 22 e 50.

BELÉM, M. d. O.; ODA, J. Y. Doenças inflamatórias intestinais: considerações fisiológicas e alternativas terapêuticas. *Arq. ciências saúde UNIPAR*, p. 73–79, 2015. Citado na página 15.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado na página 16.

BUCILUă, C.; CARUANA, R.; NICULESCU-MIZIL, A. Model compression. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2006. p. 535–541. Citado na página 32.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, Springer, v. 21, n. 1, p. 1–13, 2020. Citado na página 41.

CIUTI, G.; MENCIASSI, A.; DARIO, P. Capsule endoscopy: From current achievements to open challenges. *IEEE Reviews in Biomedical Engineering*, v. 4, p. 59–72, 2011. Citado 4 vezes nas páginas 16, 23, 25 e 26.

CONG, Y.; WANG, S.; LIU, J.; CAO, J.; YANG, Y.; LUO, J. Deep sparse feature selection for computer aided endoscopy diagnosis. *Pattern Recognition*, v. 48, n. 3, p. 907–917, 2015. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320314003719>. Citado 2 vezes nas páginas 19 e 22.

COX, D. R. Statistical significance tests. *British journal of clinical pharmacology*, Wiley-Blackwell, v. 14, n. 3, p. 325, 1982. Citado na página 47.

DAI, N.; GUBLER, C.; HENGSTLER, P.; MEYENBERGER, C.; BAUERFEIND, P. Improved capsule endoscopy after bowel preparation. *Gastrointestinal Endoscopy*, v. 61, n. 1, p. 28–31, 2005. ISSN 0016-5107. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016510704024447>. Citado na página 26.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado na página 16.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Citado 4 vezes nas páginas 16, 21, 30 e 31.

ELIAKIM, R.; FIREMAN, Z.; GRALNEK, I.; YASSIN, K.; WATERMAN, M.; KOPELMAN, Y.; LACHTER, J.; KOSLOWSKY, B.; ADLER, S. Evaluation of the pillcam colon capsule in the detection of colonic pathology: results of the first multicenter, prospective, comparative study. *Endoscopy*, © Georg Thieme Verlag KG Stuttgart · New York, v. 38, n. 10, p. 963–970, 2006. Citado na página 24.

EWALD, T. A.; FRISANCO, M. G.; ROMANINI, J. R.; MARTINIANO, L. C.; PEREIRA, S. F.; SILVA, A. C. da; NETO, J. D. da S.; SILVA, A. M. C. da; SHIMOYA-BITTENCOURT, W. Tendência temporal de mortalidade por doenças do trato gastrointestinal. *COORTE-Revista Científica do Hospital Santa Rosa*, n. 12, 2021. Citado na página 15.

GJESTANG, H. L.; HICKS, S. A.; THAMBAWITA, V.; HALVORSEN, P.; RIEGLER, M. A. A self-learning teacher-student framework for gastrointestinal image classification. In: IEEE. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.], 2021. p. 539–544. Citado na página 50.

GOSSUM, A. V.; MUNOZ-NAVAS, M.; FERNANDEZ-URIEN, I.; CARRETERO, C.; GAY, G.; DELVAUX, M.; LAPALUS, M. G.; PONCHON, T.; NEUHAUS, H.; PHILIPPER, M. et al. Capsule endoscopy versus colonoscopy for the detection of polyps and cancer. *New England Journal of Medicine*, Mass Medical Soc, v. 361, n. 3, p. 264–270, 2009. Citado 2 vezes nas páginas 25 e 26.

GOU, J.; YU, B.; MAYBANK, S. J.; TAO, D. Knowledge distillation: A survey. *International Journal of Computer Vision*, Springer, v. 129, n. 6, p. 1789–1819, 2021. Citado na página 32.

HAN, K.; WANG, Y.; CHEN, H.; CHEN, X.; GUO, J.; LIU, Z.; TANG, Y.; XIAO, A.; XU, C.; XU, Y. et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, 2022. Citado 4 vezes nas páginas 27, 28, 29 e 31.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. Citado na página 45.

HINTON, G.; VINYALS, O.; DEAN, J. et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, v. 2, n. 7, 2015. Citado na página 32.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708. Citado na página 45.

HWANG, S. Bag-of-visual-words approach to abnormal image detection in wireless capsule endoscopy videos. In: SPRINGER. *International Symposium on Visual Computing*. [S.l.], 2011. p. 320–327. Citado na página 25.

IAKOVIDIS, D. K.; GEORGAKOPOULOS, S. V.; VASILAKAKIS, M.; KOULAOUZIDIS, A.; PLAGIANAKOS, V. P. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Transactions on Medical Imaging*, v. 37, n. 10, p. 2196–2210, 2018. Citado na página 15.

JIA, X.; MENG, M. Q.-H. Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and cnn features. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.: s.n.], 2017. p. 3154–3157. Citado 2 vezes nas páginas 20 e 22.

JIA, X.; XING, X.; YUAN, Y.; XING, L.; MENG, M. Q.-H. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE*, v. 108, n. 1, p. 178–197, 2020. Citado 2 vezes nas páginas 23 e 24.

KHAN, S.; NASEER, M.; HAYAT, M.; ZAMIR, S. W.; KHAN, F. S.; SHAH, M. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, 2021. Citado 2 vezes nas páginas 27 e 28.

KHORSHIDI, M.; DJAFARIAN, K.; AGHAYEI, E.; SHAB-BIDAR, S. A posteriori dietary patterns and risk of inflammatory bowel disease: a meta-analysis of observational studies. *International Journal for Vitamin and Nutrition Research*, Hogrefe Verlag, 2019. Citado na página 15.

KLANG, E.; BARASH, Y.; MARGALIT, R. Y.; SOFFER, S.; SHIMON, O.; ALBSHESH, A.; BEN-HORIN, S.; AMITAI, M. M.; ELIAKIM, R.; KOPYLOV, U. Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy. *Gastrointestinal endoscopy*, Elsevier, v. 91, n. 3, p. 606–613, 2020. Citado 2 vezes nas páginas 20 e 22.

KRÖNER, P. T.; ENGELS, M. M.; GLICKSBERG, B. S.; JOHNSON, K. W.; MZAIK, O.; HOOFT, J. E. van; WALLACE, M. B.; EL-SERAG, H. B.; KRITTANAWONG, C. Artificial intelligence in gastroenterology: A state-of-the-art review. *World journal of gastroenterology*, Baishideng Publishing Group Inc, v. 27, n. 40, p. 6794, 2021. Citado na página 19.

KWACK, W. G.; LIM, Y. J. Current status and research into overcoming limitations of capsule endoscopy. *Clinical endoscopy*, The Korean Society of Gastrointestinal Endoscopy, v. 49, n. 1, p. 8–15, 2016. Citado na página 26.

LI, P.; LI, Z.; GAO, F.; WAN, L.; YU, J. Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images. In: IEEE. *2017 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.], 2017. p. 1518–1523. Citado 2 vezes nas páginas 20 e 22.

LIMA, D. L. S.; PESSOA, A. C. P.; PAIVA, A. C. D.; CUNHA, A. M. Trigueiros da S.; JúNIOR, G. B.; ALMEIDA, J. D. S. D. Classification of video capsule endoscopy images using visual transformers. In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. [S.l.: s.n.], 2022. p. 1–4. Citado na página 52.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. V. D.; GINNEKEN, B. V.; SÁNCHEZ, C. I. A survey

on deep learning in medical image analysis. *Medical image analysis*, Elsevier, v. 42, p. 60–88, 2017. Citado 3 vezes nas páginas 15, 16 e 20.

MARSLAND, S. *Machine Learning: An Algorithmic Perspective, Second Edition*. 2nd. ed. [S.l.]: Chapman & Hall/CRC, 2014. ISBN 1466583282. Citado na página 40.

MURUGANANTHAM, P.; BALAKRISHNAN, S. M. A survey on deep learning models for wireless capsule endoscopy image analysis. *International Journal of Cognitive Computing in Engineering*, Elsevier, v. 2, p. 83–92, 2021. Citado na página 45.

MURUGANANTHAM, P.; BALAKRISHNAN, S. M. Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization. *Journal of Medical and Biological Engineering*, Springer, v. 42, n. 2, p. 157–168, 2022. Citado 4 vezes nas páginas 21, 22, 49 e 50.

NAWARATHNA, R.; OH, J.; MUTHUKUDAGE, J.; TAVANAPONG, W.; WONG, J.; GROEN, P. C. D.; TANG, S. J. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *Neurocomputing*, Elsevier, v. 144, p. 70–91, 2014. Citado 2 vezes nas páginas 19 e 22.

RADOSAVOVIC, I.; KOSARAJU, R. P.; GIRSHICK, R.; HE, K.; DOLLÁR, P. Designing network design spaces. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 10428–10436. Citado na página 33.

RONDONOTTI, E.; HERRERIAS, J. M.; PENNAZIO, M.; CAUNEDO, A.; MASCARENHAS-SARAIVA, M.; FRANCHIS, R. de. Complications, limitations, and failures of capsule endoscopy: a review of 733 cases. *Gastrointestinal endoscopy*, Elsevier, v. 62, n. 5, p. 712–716, 2005. Citado na página 24.

SAITO, H.; AOKI, T.; AOYAMA, K.; KATO, Y.; TSUBOI, A.; YAMADA, A.; FUJISHIRO, M.; OKA, S.; ISHIHARA, S.; MATSUDA, T. et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal endoscopy*, Elsevier, v. 92, n. 1, p. 144–151, 2020. Citado 2 vezes nas páginas 20 e 22.

SHAMSHAD, F.; KHAN, S.; ZAMIR, S. W.; KHAN, M. H.; HAYAT, M.; KHAN, F. S.; FU, H. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022. Citado na página 21.

SHEN, Z.; BELLO, I.; VEMULAPALLI, R.; JIA, X.; CHEN, C.-H. Global self-attention networks for image recognition. *arXiv preprint arXiv:2010.03019*, 2020. Citado na página 16.

SMEDSRUD, P. H.; THAMBAWITA, V.; HICKS, S. A.; GJESTANG, H.; NEDREJORD, O. O.; NÆSS, E.; BORGLI, H.; JHA, D.; BERSTAD, T. J. D.; ESKELAND, S. L. et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, Nature Publishing Group, v. 8, n. 1, p. 1–10, 2021. Citado 4 vezes nas páginas 21, 35, 36 e 37.

SRIVASTAVA, A.; TOMAR, N. K.; BAGCI, U.; JHA, D. Video capsule endoscopy classification using focal modulation guided convolutional neural network. In: IEEE. *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.], 2022. p. 323–328. Citado 4 vezes nas páginas 21, 22, 49 e 50.

SUN, C.; SHRIVASTAVA, A.; SINGH, S.; GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 843–852. Citado na página 31.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 3, p. 209–249, 2021. Citado na página 15.

TOUVRON, H.; CORD, M.; DOUZE, M.; MASSA, F.; SABLAYROLLES, A.; JéGOU, H. *Training data-efficient image transformers & distillation through attention*. 2021. Citado 2 vezes nas páginas 32 e 33.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado 4 vezes nas páginas 27, 28, 29 e 31.

WANG, A.; BANERJEE, S.; BARTH, B. A.; BHAT, Y. M.; CHAUHAN, S.; GOTTLIEB, K. T.; KONDA, V.; MAPLE, J. T.; MURAD, F.; PFAU, P. R. et al. Wireless capsule endoscopy. *Gastrointestinal endoscopy*, Elsevier, v. 78, n. 6, p. 805–815, 2013. Citado 3 vezes nas páginas 24, 25 e 26.

YUAN, Y.; LI, B.; MENG, M. Q.-H. Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images. *IEEE Transactions on Automation Science and Engineering*, v. 13, n. 2, p. 529–535, 2016. Citado 2 vezes nas páginas 19 e 22.