Classification of Lesions in Capsule Endoscopy Images using Swin Transformer and Semi-Supervised Learning

Alejandro Costa de Oliveira[1], Mario Vítor Vieira Cella[1], Darlan Bruno Pontes Quintanilha[1], Celso Luiz Silva Soares Filho[1], Francisco Glaubos Nunes Clímaco[1], Tiago Bonini Borchartt[1], Anselmo Cardoso de Paiva[1]

[1] Núcleo de Computação Aplicada, Universidade Federal do Maranhão (UFMA)
Caixa Postal 65085-580, São Luís, MA, Brasil

{alejandro.costa, mario.cella, dquintanilha, celso, paiva}@nca.ufma.br

{francisco.glaubos, tiago.bonini}@ufma.br

Abstract. *Automated analysis of images obtained by Wireless Capsule Endoscopy (WCE) is a significant challenge in the medical field, especially due to the difficulty in lesion detection, the scarcity of labeled samples, and the high visual variability of the images. This work proposes a multiclass classification method for luminal findings in WCE images based on the Swin Transformer architecture, structured in two sequential stages: a binary classifier, responsible for filtering normal from anomalous images, followed by a multiclass classifier for identifying the specific lesion. To address the limitation of labeled data, offline data augmentation and semi-supervised learning techniques were employed. Experiments performed on the Kvasir-Capsule array with six classes of luminal findings demonstrated that Transformer-based architectures consistently outperform traditional CNN models such as ResNet-50, MobileNetV3, and EfficientNetV2. The Swin Transformer model achieved 98% accuracy and an F1-score in the binary step and 86% in multiclass classification, representing a gain of 4 percentage points compared to purely supervised training.*

1. Introduction

The diagnosis of gastrointestinal tract pathologies remains a significant clinical challenge, particularly in regions that are difficult to access via conventional endoscopy, such as the small intestine, thereby complicating the detection of lesions in their early stages [Gounella et al., 2023]. Consequently, technological advancements have enabled the development of methods designed to address this issue, most notably wireless capsule endoscopy [Pan and Wang, 2012].

Wireless Capsule Endoscopy (WCE) is a non-invasive technology that enables the sequential capture of images throughout the gastrointestinal tract via a capsule ingested by the patient [Wang et al., 2013]. Unlike conventional endoscopy, WCE provides a detailed view of intestinal regions that are otherwise difficult to access. The high volume and continuity of the resulting images have established WCE as a valuable clinical tool for the detection of hemorrhages, ulcers, polyps, inflammatory processes, and lesions [Pennazio et al., 2015].

However, the analysis of WCE imagery presents significant practical and diagnostic challenges. First, the extensive volume of frames per examination, typically numbering in the tens of thousands, renders manual review extremely time-consuming and

susceptible to visual fatigue [Beg et al., 2021], thereby increasing the risk of overlooking critical images. Second, the images exhibit high variability in terms of illumination, contrast, and the presence of bubbles, food residues, capsule motility, and image artifacts, all of which complicate the detection and classification of lesions [Pan and Wang, 2012]. Finally, inter-observer variability exists in the interpretation of findings, both in detection and classification, which hinders the standardization of diagnostic outcomes [Cortegoso Valdivia et al.].

To mitigate these issues, we propose a multiclass classification method for luminal findings, such as angiectasia, erosion, erythema, lymphangiectasia, and ulcers, in WCE imagery based on Transformer architectures. This approach leverages the attention mechanisms of these networks to capture the global context of the images, combining them with semi-supervised learning that utilizes unlabeled images to expand the training set and increase the volume of available samples, particularly for rare and imbalanced classes.

Swin Transformer and Vision Transformer (ViT) models are evaluated in comparison with commonly applied CNN architectures, with the objective of identifying the approach that best exploits the global context of WCE images within an imbalanced data scenario.

2. Related works

This section presents related works, encompassing both the models utilized for WCE image classification and data augmentation methods.

CNN architectures are extensively tested in medical image classification tasks. Research in other healthcare domains, such as dental lesion analysis, has reported an accuracy of 92.10% using ResNet50 [Ferreira et al., 2023]. In the specific context of capsule endoscopy classification, CNNs also yield significant results. Studies evaluating the performance of ResNet50 report average accuracies of 98% [Li et al., 2024], while modified versions of the same architecture have achieved results of 99% [Cambay et al., 2024]. Additionally, lightweight architectures such as MobileNetV2 demonstrate high efficiency regarding the trade-off between inference time and computational cost, reaching a recorded accuracy of 91% [Kaur and Kumar, 2024].

Although traditional CNN architectures achieve robust results, recent studies have begun to analyze the inherent differences between CNNs and Transformer-based models [Ribeiro and von Wangenheim, 2025]. In WCE image classification tasks, CNNs exhibit limitations in correlating the global context of the image, given that these networks focus primarily on local features. In contrast, Transformer-based architectures utilize self-attention mechanisms to analyze the global context simultaneously. Several studies apply Transformer models to WCE image classification tasks. The Vision Transformer (ViT) is a widely utilized model for this type of activity, and conducted tests demonstrate a weighted F1-Score of 71%, a result that surpasses several traditional CNN architectures [Regmi et al., 2023].

In comparison with other CNN architectures, tests conducted with the Swin Transformer in multiclass classification scenarios demonstrate superior results relative to alternative frameworks [Bai et al., 2022]. Recent studies have achieved robust performance by

utilizing Transformer models for the classification of capsule endoscopy images. Models such as the Swin Transformer yielded high performance across evaluated metrics on the validation set, achieving 90% accuracy and a 91% F1-Score [Choudhary et al., 2024].

One of the most significant challenges in WCE image classification is class imbalance and the paucity of examples for certain pathologies, which hinders the model's ability to effectively learn rare patterns. In such instances, traditional data augmentation techniques, such as image rotation and contrast adjustment, are recommended in an effort to mitigate this issue [Kim and Lim, 2021]. However, lesion generation techniques are currently being developed to support the resolution of this problem and ensure superior results.

Adrian B. et al. (2024) propose two methods for generating realistic lesions in WCE imagery, one based on Poisson Image Editing and the other utilizing an Image Inpainting Generative Adversarial Network (GAN). The implementation of a combined approach using both techniques for synthetic image generation resulted in a performance increase of over seven percentage points compared to previous results [Chłopowiec et al., 2024], thereby demonstrating a robust method for data augmentation.

Another strategy to mitigate this problem is the adoption of unlabeled imagery for model training. Recent studies have introduced modified semi-supervised learning frameworks for WCE images, achieving an accuracy of 93,17% [Guo and Yuan, 2020].

3. Materials and Methods

This section presents the dataset utilized and the proposed method for the classification of findings in WCE images. The methodology consists of training two Swin Transformer models: a binary classifier, responsible for filtering normal from anomalous images, followed by a multiclass classifier for the specific identification of lesions within the anomalous images. Figure 1 illustrates the complete representation of the proposed method.

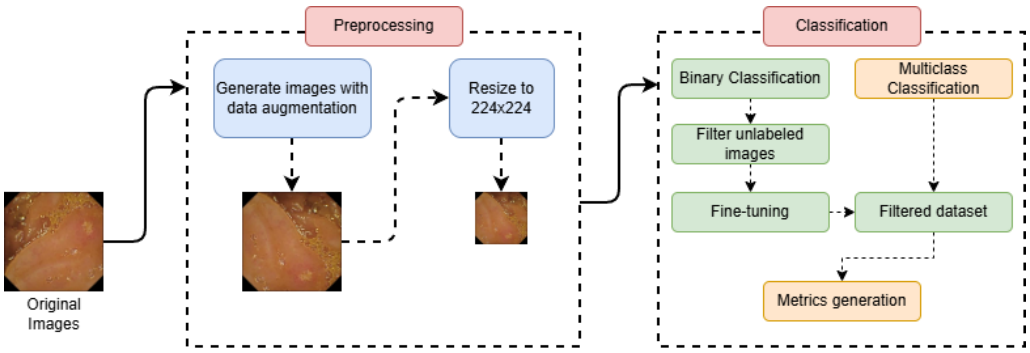

Figure 1. Proposed Method

3.1. Dataset

To conduct the proposed experiments, this study utilized the capsule endoscopy dataset provided by Kvasir-Capsule [Smedsrud et al., 2021]. The dataset comprises 117 videos that allow for the extraction of 4,741,504 frames; however, it contains only 47,238 labeled frames distributed across 14 distinct classes, as illustrated in Figure 2.

It is important to emphasize that a considerable imbalance exists between the classes, as shown in Figure 3, particularly when the objective is the multiclass classification of pathological findings within the images. This is due to the fact that the majority of classes possess no more than one thousand labeled examples available for the training of classification agents.

This necessitated the removal of certain available classes, either due to an insufficient number of examples or because they did not align with the pathology classification task. The classes utilized in this study were Angiectasia, Erosion, Erythema, Lymphangiectasia, Normal Clean Mucosa, and Ulcer. Images belonging to anatomical categories were excluded, as the focus of the research is the classification of pathological lesions. Furthermore, the low volume of labeled images relative to the vast amount of unlabeled data encourages the adoption of semi-supervised learning techniques and data augmentation strategies.

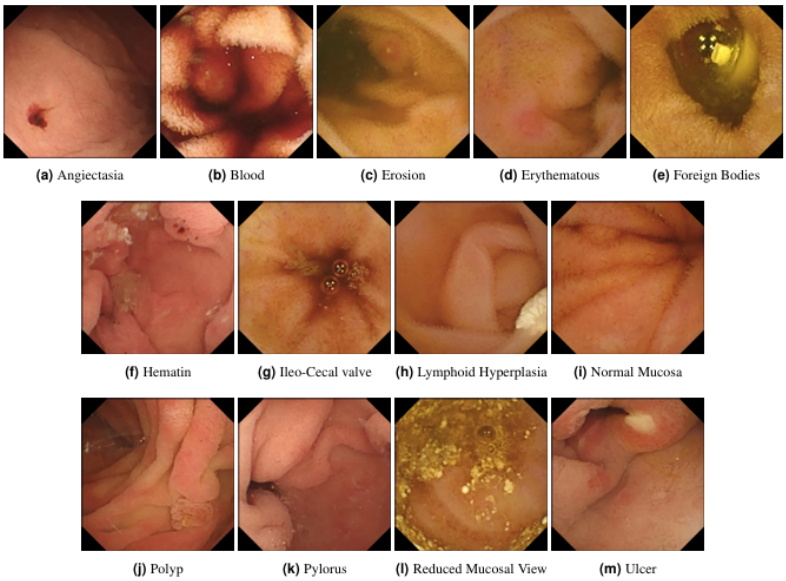

Figure 2. Examples of luminal findings available in the Kvasir-Capsule. Source: [Smedsrud et al., 2021]

3.2. Preprocessing

This subsection represents the Preprocessing stage of Figure 1.

As previously noted, a primary challenge surrounding the training of WCE classification agents arises from the scarcity of sufficient imagery to learn the patterns of specific pathologies. In an effort to overcome this obstacle, offline data augmentation was employed. Unlike online data augmentation, the offline method generates images prior to training and incorporates them as part of the training dataset. This approach is instrumental in generating additional image examples and increasing the sample size for each class.

Minor adjustments were made to image orientation and contrast in order to generate new samples while ensuring the complete representation of the pathology remains intact. The utilization of offline data augmentation facilitates greater experimental re-

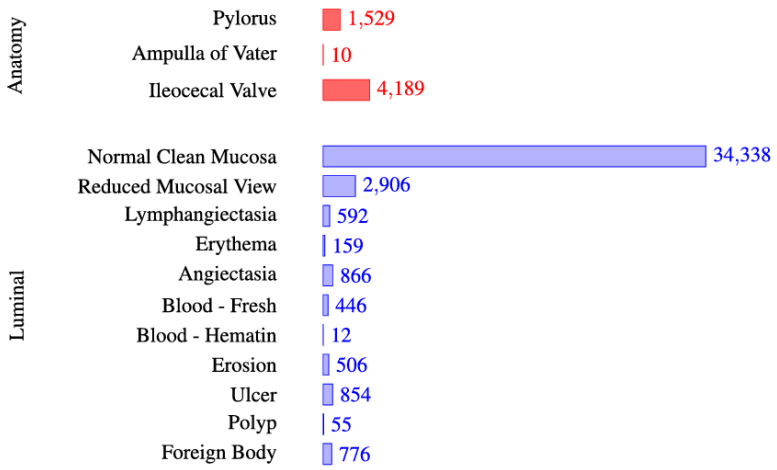

Figure 3. Distribution of anatomical landmark classes (in red) and luminal findings (in blue). Source: [Smedsrud et al., 2021]

producibility, as the dataset remains fixed throughout the training cycles [Sugimura and Hartl, 2018].

During the training phase, each image undergoes a fixed resizing process to a resolution of 224x224 pixels, with experiments conducted using batch sizes of 16 and 32. The images are converted into tensors with values normalized to a range between 0 and 1.

3.3. Classification

This subsection represents the Classification stage of Figure 1.

For the classification stage, several architectures were evaluated, specifically EfficientNetV2, MobileNetV3, ResNet-50, ViT, and the Swin Transformer [Tan and Le, 2021, Howard et al., 2019, He et al., 2016, Dosovitskiy et al., 2020, Liu et al., 2021]. The selected architectures have been extensively validated across various WCE image classification studies and demonstrate robust performance in this task. The decision to conduct experiments using both CNN and Transformer architectures stems from the objective to determine which paradigm yields superior results for the specific task proposed in this study.

The classification stage was structured into two sequential phases, designed to optimize the detection of findings and reduce the false positive rate induced by class imbalance. In the first phase, a binary classification model acts as the primary filter, distinguishing normal frames (healthy mucosa) from images containing potential pathological findings. Only the images classified as anomalous by this initial filter are forwarded to the second stage, where a multiclass model categorizes the specific lesion, such as Angiectasia, Erosion, or Ulcer.

3.4. Semi-supervised

Due to class imbalance and the limited number of samples, semi-supervised learning was employed in conjunction with other data augmentation techniques. Over two thousand unlabeled images were incorporated into the binary classification stage. The binary model resulting from the prior training phase generates a label probability for each image; if

the confidence of this prediction exceeds a predefined threshold, the image is assigned a positive or negative label. The selection of this specific threshold value is justified by previous studies demonstrating lower error rates compared to lower threshold levels [Sohn et al., 2020]. These pseudo-labeled images are integrated into the binary training set and are treated as new instances in the training cycle, during which the model undergoes a fine-tuning process.

4. Results and discussion

This section presents and discusses the experimental configurations and provides an evaluation of the results obtained across the stages of the developed method.

4.1. Experiment settings

The experiments were conducted on the Kaggle platform, utilizing a virtual machine equipped with a 16GB NVIDIA P100 GPU, a 4-core CPU, and 29GB of RAM. The implementations were developed using the Python programming language and the PyTorch framework. Six classes (Angiectasia, Erosion, Erythema, Lymphangiectasia, Normal Clean Mucosa, and Ulcer) were selected from the original dataset. The data was partitioned into training (70%), validation (15%), and testing (15%) sets. This split was performed on a per-video basis, ensuring that all frames from a single video were restricted to a single subset to prevent any data leakage. All architectures utilized the same data split, and data augmentation techniques were applied exclusively to the training subset.

Upon completion of the binary classifier training, images labeled as 'normal' (without findings) are removed from the original dataset, resulting in a subset composed exclusively of the five classes of luminal findings utilized in the multiclass stage. Subsequently, the trained multiclass model performs the classification of this generated dataset.

Due to the substantial volume of unlabeled data within the original dataset, the semi-supervised learning technique is applied to the models that achieve the most favorable results. Given that it is not possible to verify the classification of a specific pathology in the absence of original ground-truth labels, the semi-supervised method is utilized exclusively during the binary training stage.

4.2. Performance Metrics Assessment

The selected architectures were trained using the following hyperparameters: images were resized to a standard resolution of 224x224 pixels, and all models were trained with batch sizes of 16 and 32 for a total of 30 epochs. A learning rate of 1e-4 was employed along with the AdamW optimizer.

4.2.1. Image identification with findings (binary classification)

The results obtained for the binary stage are presented in Table 1. It can be observed that the architectures based on the Transformer paradigm outperformed the CNN-based networks. This indicates that the attention mechanisms inherent in Transformer networks, which facilitate a global understanding of the image, significantly aid in the identification of lesions.

Table 1. Results of image classification with and without findings

Methods \ Metrics	Precision	Recall	F1-Score	Accuracy
EfficientNetV2				
32 batch size	0.85	0.84	0.85	0.84
16 batch size	0.85	0.85	0.85	0.85
MobileNet				
32 batch size	0.90	0.90	0.90	0.90
16 batch size	0.90	0.90	0.90	0.90
ResNet50				
32 batch size	0.86	0.86	0.86	0.86
16 batch size	0.89	0.89	0.89	0.89
Swin Transformer				
32 batch size	0.94	0.93	0.93	0.93
16 batch size	**0.98**	**0.98**	**0.98**	**0.98**
ViT				
32 batch size	0.94	0.94	0.94	0.94
16 batch size	0.93	0.93	0.93	0.93

Based on these results, the Swin Transformer and ViT architectures were selected for the semi-supervised learning process due to their superior performance. In this stage, a set of over two thousand unlabeled images is incorporated exclusively into the binary classification phase. The confidence threshold was established at 95%, and the training configurations for this stage remain consistent with those previously described. The results obtained from this process are presented in Table 2.

Table 2. Results of image classification with and without findings using semi-supervised learning

Methods \ Metrics	Precision	Recall	F1-Score	Accuracy
Swin Transformer				
32 batch size	**0.96**	**0.96**	**0.96**	**0.96**
16 batch size	0.92	0.92	0.92	0.92
ViT				
32 batch size	0.93	0.92	0.92	0.92
16 batch size	0.92	0.91	0.91	0.91

It can be observed that the results in this stage, while still satisfactory, did not surpass the metrics achieved through purely supervised training. Given the high confidence threshold, the probability of images being incorrectly assigned pseudo-labels is low, yet it persists. Consequently, the fine-tuning process may reinforce certain errors, resulting in a performance decline, albeit a marginal one.

4.2.2. Classification of luminal findings (multiclass classification)

To provide a more comprehensive overview of the overall experimental results, all architectures were evaluated during the multiclass training stage. The training parameters remained identical to those previously described. Only images containing pathologies were

utilized, totaling five detectable classes: Angiectasia, Erosion, Erythema, Lymphangiectasia, and Ulcer. Consequently, the results obtained from this phase are presented in Table 3.

Table 3. Results of multiclass classification using a dataset generated by binary training

Methods \ Metrics	Precision	Recall	F1-Score	Accuracy
EfficientNetV2				
32 batch size	0.77	0.78	0.77	0.78
16 batch size	0.80	0.80	0.79	0.80
MobileNet				
32 batch size	0.79	0.79	0.79	0.79
16 batch size	0.82	0.81	0.81	0.81
ResNet50				
32 batch size	0.64	0.65	0.58	0.65
16 batch size	0.81	0.81	0.81	0.81
Swin Transformer				
32 batch size	0.82	0.82	0.82	0.82
16 batch size	**0.87**	**0.86**	**0.86**	**0.86**
ViT				
32 batch size	0.86	0.84	0.85	0.84
16 batch size	0.83	0.82	0.82	0.82

Once again, the Transformer-based networks outperformed the CNN architectures, reinforcing the premise that the attention mechanisms inherent in these designs provide significant improvements in the identification of pathological findings. Table 4 presents the results obtained from the semi-supervised binary dataset. Since CNNs were not selected for this methodological stage due to their inferior performance, only the Transformer architectures were evaluated within this specific scenario.

Table 4. Results of multiclass classification using a dataset generated by binary training using semi-supervised learning

Methods \ Metrics	Precision	Recall	F1-Score	Accuracy
Swin Transformer				
32 batch size	**0.87**	**0.86**	**0.86**	**0.86**
16 batch size	0.83	0.82	0.83	0.82
ViT				
32 batch size	0.85	0.84	0.83	0.84
16 batch size	0.84	0.84	0.83	0.84

To provide a more granular perspective of the results, Table 5 details the individual metrics for each class using the Swin Transformer model with a batch size of 32, under the semi-supervised learning framework. Subsequently, to evaluate the effectiveness of the proposed two-stage pipeline, Table 6 presents the results of the same model trained in a single-stage approach for the direct classification of all six classes (comprising the five luminal finding categories and the normal class). The comparison demonstrates that the

implementation of the proposed two-stage method yielded superior results compared to direct training without these refinements.

Table 5. Detailed metrics for Swin Transformer with normal filtering using semi-supervised learning

Classes \ Metrics	Precision	Recall	F1-Score
Angiectasia	0.96	0.87	0.91
Erosion	0.69	0.61	0.65
Erythema	0.88	0.77	0.82
Lymphangiectasia	0.89	0.91	0.90
Ulcer	0.84	0.96	0.90
Average	0.87	0.86	0.86
Accuracy		0.86	

Table 6. Detailed metrics for Swin Transformer without additional methods

Classes \ Metrics	Precision	Recall	F1-Score
Angiectasia	1.00	0.22	0.36
Erosion	0.36	0.26	0.30
Erythema	0.00	0.00	0.00
Lymphangiectasia	0.00	0.00	0.00
Normal	0.91	0.97	0.94
Ulcer	0.31	0.64	0.42
Average	0.67	0.67	0.64
Accuracy		0.67	

4.3. Discussion

Following the completion of the experiments, it is evident that architectures based on the Transformer paradigm achieved superior performance in the detection and classification of WCE imagery when compared to traditional CNNs. This can be attributed to the self-attention mechanisms inherent in Transformer architectures, which facilitate the modeling of global relationships between disparate regions of an image from the initial layers. During supervised learning, the 98% accuracy achieved in binary classification demonstrates a significant competitive advantage for the Swin Transformer, even when measured against other widely utilized architectures.

The methods employed enhanced the diversity of the imagery and ensured F1-score and accuracy results of 86% for both metrics, utilizing the Swin Transformer with a batch size of 32 in the multiclass classification stage. When compared to the traditional method without the application of semi-supervised learning, the proposed approach yields a 4% improvement; furthermore, in comparison to similar studies, the Swin architecture achieved accuracy results nearly 8% higher [Bai et al., 2022]. These gains are even more pronounced when compared to the traditional training method without the normal frame filter, which caused a considerable decline in the individual metrics across all classes. The presence of normal images in that scenario hindered the model's ability to learn the specific features and particularities of each pathological class.

The ViT architecture demonstrated highly satisfactory results when compared to the Swin Transformer. The application of semi-supervised learning maintained a consistent accuracy of 84% with a batch size of 32 compared to the purely supervised experiment; however, with a batch size of 16, the proposed method yielded a 2% improvement. Furthermore, compared to previous literature, the ViT achieved an increase of nearly 5% in the multiclass classification task [Bai et al., 2022].

Nonetheless, despite these highly positive overall results, the class-wise performance reveals significant disparities. Specifically, the Erosion class, although represented by a substantial number of images, often contains lesions that are minute and difficult to discern. In contrast, classes such as Ulcer and Lymphangiectasia present more prominent features that are more readily identifiable, consequently yielding some of the highest performance metrics.

The distinctive feature of the proposed method is the reduction in the number of images a specialist needs to analyze during the examination, since the trained binary model acts as a filter for healthy images, and the multiclass model evaluates only images with luminal findings. In this way, the pipeline can be implemented in the workflow with medical CAD tools that integrate AI agents during the examination to perform image classification.

5. Conclusion

Considering the comprehensive set of results obtained, it is concluded that the method proposed in this study was effective for the classification of Wireless Capsule Endoscopy (WCE) images from the Kvasir Capsule dataset. The tested models achieved performance peaks of 98% in accuracy and F1-score for binary classification, and 86% for both metrics in the multiclass stage. These results demonstrate that the integration of data augmentation techniques and semi-supervised learning methods can significantly enhance performance in similar medical imaging tasks.

Regarding future work, alternative methods for incorporating unlabeled images within the multiclass training stage are expected to be explored. The generation of synthetic lesions is a significant possibility, given the performance gains such methods can provide, although they are considerably more complex than the traditional data augmentation techniques employed in this study. Furthermore, expanding the methodology toward full-video classification could help preserve temporal information from frames adjacent to the analyzed image; this contextual data can be leveraged to achieve more accurate and reliable predictions.

6. Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Fundação de Amparo a Pesquisa do Maranhão (FAPEMA), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Empresa Brasileira de Serviços Hospitalares (Ebserh) Brazil (Proc. 409593/2021-4).

References

Long Bai, Liangyu Wang, Tong Chen, Yuanhao Zhao, and Hongliang Ren. Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset. *Electronics*, 11(17):2747, 2022.

Sabina Beg, Tim Card, Reena Sidhu, Ewa Wronska, Krish Ragunath, Hey-Long Ching, Anastasios Koulaouzidis, Diana Yung, Simon Panter, Mark Mcalindon, et al. The impact of reader fatigue on the accuracy of capsule endoscopy interpretation. *Digestive and Liver Disease*, 53(8):1028–1033, 2021.

Veysel Yusuf Cambay, Prabal Datta Barua, Abdul Hafeez Baig, Sengul Dogan, Mehmet Baygin, Turker Tuncer, and UR Acharya. Automated detection of gastrointestinal diseases using resnet50*-based explainable deep feature engineering model with endoscopy images. *Sensors*, 24(23):7710, 2024.

Adrian B Chłopowiec, Adam R Chłopowiec, Krzysztof Galus, Wojciech Cebula, and Martin Tabakov. Local lesion generation is effective for capsule endoscopy image data augmentation in a limited data setting. *arXiv preprint arXiv:2411.03098*, 2024.

Abhishek Choudhary, Mayur Raj, and Kanishk Kumar. High-performance capsule endoscopy classification using swin transformers, 2024.

P Cortegoso Valdivia, U Deding, T Bjørsum-Meyer, G Baatrup, I Fernández-Urién, X Dray, P Boal-Carvalho, P Ellul, E Toth, E Rondonotti, et al. Inter/intra-observer agreement in video-capsule endoscopy: Are we getting it all wrong? a systematic review and meta-analysis. diagnostics. 2022; 12 (10): 2400.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mateus F de C Ferreira, Paula D Portella, Juliana F de Souza, Bruna C Dias, Luciana R da S Assunção, and Lucas F de Oliveira. Avaliaçao do uso redes neurais convolucionais para identificaçao de lesoes cariosas dentárias. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 473–478. SBC, 2023.

Rodrigo Gounella, Talita Conte Granado, Oswaldo Hideo Ando Junior, Daniel Luís Luporini, Mario Gazziro, and João Paulo Carmo. Endoscope capsules: The present situation and future outlooks. *Bioengineering*, 10(12):1347, 2023.

Xiaoqing Guo and Yixuan Yuan. Semi-supervised wce image classification with adaptive aggregated attention. *Medical Image Analysis*, 64:101733, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

Parminder Kaur and Rakesh Kumar. Performance analysis of convolutional neural network architectures over wireless capsule endoscopy dataset. *Bulletin of Electrical Engineering and Informatics*, 13(1):312–319, 2024.

Sang Hoon Kim and Yun Jeong Lim. Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges. *Diagnostics*, 11(9):1722, 2021.

Dongguang Li, David Cave, April Li, and Shaoguang Li. Enhanced accuracy for classification of video capsule endoscopy images using multiple deep learning convolutional neural networks. *iGIE*, 3(1):72–81, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Guobing Pan and Litong Wang. Swallowable wireless capsule endoscopy: Progress and technical challenges. *Gastroenterology research and practice*, 2012(1):841691, 2012.

Marco Pennazio, Cristiano Spada, Rami Eliakim, Martin Keuchel, Andrea May, Chris J Mulder, Emanuele Rondonotti, Samuel N Adler, Joerg Albert, Peter Baltes, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European society of gastrointestinal endoscopy (esge) clinical guideline. *Endoscopy*, 47(04):352–386, 2015.

Smriti Regmi, Aliza Subedi, Ulas Bagci, and Debesh Jha. Vision transformer for efficient chest x-ray and gastrointestinal image classification, 2023. URL `https://arxiv.org/abs/2304.11529`.

Rodrigo PS Ribeiro and Aldo von Wangenheim. Instance segmentation in medical imaging: A comparative study of cnn and transformer-based models in a teledermatology study-case. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 819–827. SBC, 2025.

Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8 (1):142, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Peter Sugimura and Florian Hartl. Building a reproducible machine learning pipeline. *arXiv preprint arXiv:1810.04570*, 2018.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

Amy Wang, Subhas Banerjee, Bradley A Barth, Yasser M Bhat, Shailendra Chauhan, Klaus T Gottlieb, Vani Konda, John T Maple, Faris Murad, Patrick R Pfau, et al. Wireless capsule endoscopy. *Gastrointestinal endoscopy*, 78(6):805–815, 2013.