**Thick Data Analytics for Detecting Inflammatory Bowel Diseases (IBD) Based on Vision Transformers**

**by**  
**[Abel Serracín]**

[Bachelor of Engineering in Computer Systems, 2021, Universidad Latina de Panamá]

[Project] Submitted in Partial Fulfillment of the  
Requirements for the Degree of  
[Master of Computer Science]

in the  
[Department of Computer Science]  
[Faculty of Science and Environmental Studies]

© [Abel, Serracín] 2025  
LAKEHEAD UNIVERSITY  
[WINTER 2025]

Copyright in this work is held by the author. Please ensure that any reproduction   
or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

|  |  |  |
| --- | --- | --- |
| **Name:** | **[Firstname Lastname]** | |
| **Degree:** | **[Master of …/Doctor of …]** | |
| **Title:** | **Type or paste the title of the work here** | |
| **Committee:** | **Chair:** | **[Firstname Lastname]** [Academic Role, Academic Unit] |
|  | **[Firstname Lastname]** Supervisor [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** Committee Member [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** Committee Member [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** Committee Member [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** Committee Member [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** Examiner [Academic Role, Academic Unit] | |
|  | **[Firstname Lastname]** External Examiner  [Academic or Professional Role, Academic Unit] [Institution] | |

Please review the Declaration of Committee page instructions on the library’s thesis website before completing this page:   
[**https://www.lib.sfu.ca/help/publish/thesis/format/declaration-committee**](https://www.lib.sfu.ca/help/publish/thesis/format/declaration-committee)

Remember to delete this note before submission.

Ethics Statement



This page must appear as p. iii if your research required ethics approval. Remember to delete this note before submission.

If an ethics exemption was granted, this page must be removed.

If ethics approval was not required, this page must be removed.

Remember to update the Table of Contents after removing this page.

Abstract

Abstract paragraphs should be unindented: use the style **1\_Para\_FlushLeft.** Master’s abstracts are limited to 150 words; the limit is 350 words for doctoral abstracts. Abstract text must fit on a single page. Keywords may appear alone on a second page.

You may include up to six keywords or phrases. Keywords should be separated with semicolons.

**Keywords**: thesis template; Microsoft Word; keywords; 1” hanging indent; instructions

Dedication

This is an optional page. Use your choice of paragraph style for text on this page (**1\_Para\_FlushLeft** shown here).

To hide the heading at the top of this page, select the text and change the text colour to white.

Acknowledgements

This is an optional page. Use your choice of paragraph style for text on this page (**1\_Para** shown here).

Table of Contents

[Declaration of Committee ii](#_Toc66285129)

[Ethics Statement iii](#_Toc66285130)

[Abstract iv](#_Toc66285131)

[Dedication v](#_Toc66285132)

[Acknowledgements vi](#_Toc66285133)

[Table of Contents vii](#_Toc66285134)

[List of Tables viii](#_Toc66285135)

[List of Figures ix](#_Toc66285136)

[List of Acronyms x](#_Toc66285137)

[Glossary xi](#_Toc66285138)

[Preface/Executive Summary/Image xii](#_Toc66285139)

[Chapter 1. Introduction 1](#_Toc66285140)

[References 2](#_Toc66285141)

[Appendix A. An Example of an Appendix 3](#_Toc66285142)

List of Tables

**No table of figures entries found.**

Use References>Insert Caption to create caption labels and numbers. Right-click on the text above and select Update Field to update this list. See the Thesis Template Instructions for information on creating tables, figures, and captions.

Remember to delete this note before submission.

If there are no tables in the document, remove this page.

List of Figures

**No table of figures entries found.**

Use References>Insert Caption to create caption labels and numbers. Right-click on the text above and select Update Field to update this list. See the Thesis Template Instructions for information on creating tables, figures, and captions.

Remember to delete this note before submission.

If there are no figures in the document, remove this page.

List of Acronyms

|  |  |
| --- | --- |
| SFU | Simon Fraser University |
| LAC | Library and Archives Canada |
|  | This is an optional page and can be removed if not used. |
|  | Use one table row for each item to allow sorting using Word’s table tools. |
|  | Apply the style **1\_Para\_NoSpace** to table rows as shown here. |
|  |  |
|  |  |

Glossary

|  |  |
| --- | --- |
| Thesis | An extended research paper that is part of the final exam process for a graduate degree. The document may also be classified as a project or collection of extended essays. |
| Glossary | An alphabetical list of key terms |
|  | This is an optional page and can be removed if not used. |
|  | Use one table row for each item to allow sorting using Word’s table tools. |
|  | Apply the style **1\_Para\_NoSpace** to table rows as shown here. |
|  |  |

Preface/Executive Summary/Image

This page can be used for a Preface, Executive Summary, or introductory image. This is an optional page and can be deleted if not used.

To hide the heading at the top of this page – e.g., if using an introductory image – select the text and change the text colour to white.

Do not delete the section break that follows this paragraph. If the section break is not visible, turn on non-printing characters using the Show/Hide icon (¶) on the Home ribbon.

# Introduction

# Chapter 2: Vision Transformers (ViTs) and Spatial Transformers for Early IBD Identification

**Vision Transformers (ViTs)**  
ViTs divide images into patches and process them as sequences, enabling the capture of long-range spatial relationships critical for identifying subtle IBD patterns like mucosal irregularities or inflammation in endoscopic images [1]. For IBD detection, ViTs:

* Analyze endoscopic/colonoscopy images to detect early inflammation, ulcers, or abnormal tissue. For example, a 2022 study demonstrated ViTs’ effectiveness in classifying Crohn’s disease severity from endoscopic data by recognizing global patterns like ulceration [2] .
* Integrate local and global features through self-attention mechanisms, allowing them to identify complex disease signatures (e.g., differentiating active inflammation from chronic scarring) that require both fine-grained and holistic analysis [2], [3].
* Leverage self-supervised learning (e.g., DINO [4]) to train on large, unlabeled medical datasets, addressing the scarcity of annotated IBD images.

**Spatial Transformers (STNs)**  
STNs dynamically adjust input images to focus on diagnostically relevant regions, improving robustness to variations in endoscopic imaging conditions [4], [5]. For IBD diagnostics, STNs:

* Automatically localize regions of interest, such as inflamed mucosa or ulcers, even in low-quality or distorted images. This capability has been validated in histopathology, where STNs improved lesion detection accuracy by aligning tissue samples [5].
* Correct procedural variations (e.g., camera angle, scale) during colonoscopy, standardizing inputs for downstream models. A 2019 study showed STNs enhanced segmentation accuracy in medical imaging by compensating for rotational and scaling artifacts [5].

# Chapter 3. How are ViTs and Spatial Transformers different from classic neural network approaches?

**Vision Transformers vs. Convolutional Neural Networks (CNNs):**

* + **CNNs** rely on convolutional layers to extract local features from small patches of the image. They are highly effective in tasks like medical image analysis but struggle with capturing global relationships between distant image regions [6].
  + **ViTs**, on the other hand, do not rely on convolutions but treat images as a sequence of patches. This allows ViTs to better capture long-range dependencies between different parts of an image, making them more powerful for tasks where both local and global context matter, such as identifying complex, spatially dispersed symptoms of IBD [6].

**Spatial Transformers vs. Classic CNNs:**

**Spatial Transformer Networks (STNs):**

* Dynamic Spatial Manipulation: STNs introduce a learnable module that enables networks to actively rectify and align input images, addressing variations in scale, rotation, and translation. This capability is particularly beneficial in medical imaging, where patient movement and differing acquisition angles can introduce inconsistencies [5].
* Enhanced Registration Accuracy: By integrating STNs, models can achieve more precise image registration. For instance, the Image-and-Spatial Transformer Network (ISTN) framework has demonstrated improved alignment of anatomical structures, which is crucial for accurate diagnosis and treatment planning [5].

**Convolutional Neural Networks (CNNs):**

* Feature Extraction and Image Understanding: CNNs are adept at extracting hierarchical features from medical images, facilitating tasks such as disease classification, tumor detection, and organ segmentation. Their architecture, comprising convolutional layers, pooling, and activation functions, allows for the modeling of complex patterns within imaging data [7].
* Proven Clinical Applications: CNNs have been successfully applied across various medical imaging modalities, including radiography, MRI, and CT scans. They have demonstrated proficiency in identifying pathologies, segmenting anatomical structures, and even predicting disease progression, thereby supporting clinical decision-making [8].

# Chapter 4. What are the best Vision Transformers and Spatial Transformer models for detecting IBDs?

The field of medical image analysis has witnessed remarkable advancements in artificial intelligence, particularly in the application of transformer-based architecture for detecting Inflammatory Bowel Diseases (IBDs). These sophisticated models have revolutionized how we process and analyze medical imaging data, offering unprecedented accuracy and reliability in identifying subtle inflammatory patterns and disease markers. By leveraging the power of self-attention mechanisms and spatial awareness, modern vision transformers and spatial transformer models have emerged as promising tools for gastroenterologists and medical professionals in their quest to improve IBD diagnosis and monitoring. Understanding the capabilities and applications of these leading models is crucial for developing robust and reliable diagnostic systems. Let's examine some of the most effective transformer architectures currently being employed in IBD detection:

* **Vision Transformer Models:**
  + Vision Transformer Models have revolutionized medical image analysis, particularly for IBD detection. The standard Vision Transformer (ViT), introduced by Dosovitskiy et al. in their seminal paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2020), has demonstrated remarkable capabilities in medical imaging. When applied to IBD detection, ViT divides colonoscopy or biopsy images into patches, treating them as sequences that enable the model to identify subtle patterns indicative of inflammation. Recent studies have shown ViT achieving accuracy rates above 90% in detecting IBD-related mucosal changes [1].
  + The Swin Transformer, presented by Liu et al. in "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" (2021), brings a significant advancement through its hierarchical architecture. This model is particularly effective for IBD detection because it processes medical images at multiple scales, like how gastroenterologists examine both macro and microscopic features of bowel inflammation. The shifted window approach allows the model to capture both local inflammatory markers and broader disease patterns, making it especially valuable for analyzing high-resolution endoscopic images.[9].
  + DINO (Self-supervised Vision Transformer), developed by Caron et al. in "Emerging Properties in Self-Supervised Vision Transformers" (2021), represents a breakthrough in handling limited labeled medical data - a common challenge in IBD research. DINO's self-supervised learning approach enables it to learn meaningful representations from unlabeled colonoscopy images, which can then be fine-tuned with a smaller set of labeled data. This makes it particularly valuable for healthcare institutions with large repositories of unlabeled endoscopic imagery [4].
* **Spatial Transformer Models:**
  + Spatial Transformer Networks (STNs), first introduced by Jaderberg et al. in "Spatial Transformer Networks" (2015), have proven especially valuable when integrated into IBD detection systems. These networks excel at automatically learning to focus on relevant regions of medical images, effectively zooming in on areas showing signs of inflammation or ulceration while maintaining spatial awareness of the surrounding tissue context. This capability is crucial for accurate IBD staging and progression monitoring [10].
  + Deformable DETR, proposed by Zhu et al. in "Deformable DETR: Deformable Transformers for End-to-End Object Detection" (2020), represents a significant advancement in medical image analysis. When applied to IBD detection, its deformable attention mechanism enables precise localization of inflammatory regions, regardless of their shape or size. The model has shown promise in distinguishing between different types of IBD lesions, such as differentiating between Crohn's disease and ulcerative colitis manifestations [11].
  + Recent research has demonstrated significant progress in hybrid transformer approaches for medical image analysis. A notable example is the TransMed framework, introduced by Ma et al. (2021) in their paper "TransMed: Transformers Advance Multi-modal Medical Image Classification." This framework showcases the potential of combining transformer architectures with multi-modal fusion strategies for medical image classification. The TransMed approach is particularly relevant for complex medical imaging tasks, as it can process and integrate information from multiple imaging modalities simultaneously while maintaining high accuracy in detection and classification tasks [12].

# Chapter 5. How can I improve the work of vision transformers using thick data approaches:

## What if I employed Siamese Neural Networks for one-shot or multi-shots learning?

A Siamese neural network (sometimes called a twin neural network) is an artificial neural network that uses the same weights while working in tandem on two different input vectors to compute comparable output vectors. Performing training using One-Shot or Few-Shot learning means that we will utilize a pre-trained Neural network on the Most or some of the categories of the Kvasir Dataset and then utilize a control category which will then be used to perform the One or Few shot learning. In theory, it works the following way: We have a trained network the classifies Dogs and Cats, but then we want it to classify horses as well. The problem is that we have few horse images or even worse we only have a single image of a horse as our dataset. So, we utilize the model to accurately classify what a Dog or a Cat is. And then use the few examples of a Horse to recognize that it is not either of those previous 2 classes: Correctly classifying it as the other(Horse). To replicate this, we will first train a Siamese Neural Network on all classes on the Kvasir dataset except the normal “polyps” class, after trained we will introduce it in a few shots learning approach and measure model performance at different numbers of shots up to 10.

For the Siamese neural network, we first approached trying to train our own Siamese NN based on the Vision Transformer (ViT) model. This approach had to be scratched as our hardware resources severely limited the possible training/fine tuning of this model to the Kvasir-V2 dataset. We then opted for the ViTMSN model for Image Classification introduced by Assran et al in 2022, [13]. Which is a ViT base masked Siamese neural network. Their approach is a self-supervised learning framework designed for learning image representations. It works by aligning the representation of an image view that includes randomly masked patches with the representation of the original, unmasked image. This pre-training strategy is especially scalable when used with Vision Transformers because the network only processes the unmasked patches. As a result, MSNs enhance the scalability of joint-embedding architecture while generating highly semantic representations that perform well in low-shot image classification. It is A graph with a line going up

AI-generated content may be incorrect.pretrained as well, we just had to fine tune it our dataset [13].

|  |  |
| --- | --- |
| Shots | Accuracy(%) |
| 1 | 75 |
| 2 | 60 |
| 3 | 52.5 |
| 4 | 33.75 |
| 5 | 13.75 |
| 6 | 13.75 |
| 7 | 13.75 |
| 8 | 13.75 |
| 9 | 13.75 |
| 10 | 13.75 |

Table 1- Model accuracy across several shots

Figure 1- Few-Shot learning model performance

After the model got finetuned, it achieved an accuracy of 83.75% on Testing. We then proceeded to do our single and few shot experiments, which yielded fascinating and somewhat counterintuitive results. The model's performance showed a distinctive pattern in which increasing the number of training examples led to decreased accuracy, challenging common assumptions about the relationship between data quantity and model performance.

As shown in both Table 1 and Figure 1, the test accuracy starts remarkably high at 75% with One-shot learning, demonstrating the model's impressive ability to learn from minimal data. However, this performance begins to decline steadily as more shots are introduced. The accuracy drops to 60% with two shots and continues decreasing to 52.5% with three shots. By the time we reach four shots, the accuracy has fallen significantly to 33.75%. Perhaps most intriguingly, the model's performance stabilizes at 13.75% from five shots onward, maintaining this constant accuracy through ten shots.

This performance pattern is particularly noteworthy because it differs substantially from what we typically expect in machine learning, where additional training data usually improves model performance. The dramatic decline in accuracy suggests that the Siamese Neural Network architecture, while highly effective at one-shot learning, may struggle to effectively integrate information from multiple examples. This could be due to several factors, including potential overfitting to the additional examples or the introduction of conflicting patterns that interfere with the model's learned representations.

The results provide strong evidence that, at least for this Siamese Neural Network architecture, One-shot learning appears to be the most effective approach. This finding has important implications for the practical application of few-shot learning in similar contexts, suggesting that carefully curated single examples might be more valuable than larger sets of training data. The consistent plateau at 13.75% accuracy after five shots further emphasizes that simply adding more training examples does not necessarily lead to better performance, highlighting the complexity of few-shot learning dynamics in real-world applications.

## 5.2 What if I used kind of YOLO approaches to zoom into region of interests?

Vision Transformers (ViTs) have demonstrated strong performance in image recognition by leveraging self-attention mechanisms to capture global relationships within an image. However, their patch-based processing makes them less effective at detecting small objects and fine-grained details, particularly in complex visual tasks requiring contextual awareness [1].Thick data, which incorporates qualitative insights such as expert annotations and domain-specific knowledge, has been proposed to enhance ViTs by improving interpretability and guiding model attention. Yet, without an efficient method for focusing on key areas of an image, much of the thick data’s potential remains underutilized.

Integrating YOLO-based region detection into the ViT + Thick Data framework can significantly enhance its effectiveness. YOLO is a real-time object detection model that quickly identifies and localizes key objects within an image [14].By using YOLO to extract Regions of Interest (ROIs) before passing them to ViTs, the model can process only the most relevant portions of an image at higher resolution, reducing unnecessary computations while improving small object recognition. This structured filtering ensures that thick data annotations are applied only to meaningful regions, enhancing their impact by reducing noise and improving interpretability.

This combined approach—where YOLO first detects important objects, thick data provides contextual insights, and ViTs analyze refined image regions—creates a more precise and efficient vision system. For example, in medical imaging, YOLO can highlight potential abnormalities, allowing ViTs to analyze them more effectively while incorporating expert annotations to improve classification accuracy [15].By adding YOLO as a pre-processing step, the original ViT + Thick Data combination becomes more efficient, interpretable, and capable of handling real-world challenges where both quantitative and qualitative insights are essential.

Our approach leverages a Vision Transformer (ViT) model for classifying medical images from the Kvasir dataset. We integrate multiple stages into our pipeline: data preprocessing, augmentation, region of interest (ROI) detection, training, and evaluation. The goal is to enhance model performance while ensuring robustness and generalizability in medical image classification.

One of the main challenges in medical image classification is ensuring that the model focuses on the most relevant regions of an image. To address this, we implemented an ROI detection module using OpenCV’s contour detection methods. This helps isolate key features in the image, removing unnecessary background noise. We applied grayscale conversion, Gaussian blurring, and Otsu’s thresholding before detecting contours and extracting meaningful regions. By cropping images to the largest detected ROI, we ensure that the model learns from the most informative areas rather than extraneous content [16].

Medical datasets are often imbalanced and small, so we apply augmentation techniques to improve generalization. We use Albumentations to introduce horizontal/vertical flips, random rotations, brightness contrast adjustments, Gaussian noise, and blur effects. These transformations simulate variations in real-world data and improve model robustness. After augmentation, we use the Hugging Face ViTImageProcessor to preprocess images into the format required for Vision Transformers [17].

For classification, we use a pre-trained Vision Transformer (ViT) from Hugging Face. ViT has demonstrated strong performance in medical imaging tasks due to its ability to model long-range dependencies in an image. We fine-tune the model on our dataset using a cross-entropy loss function and the AdamW optimizer. A key aspect of our training strategy is early stopping, which helps prevent overfitting by monitoring validation loss and halting training when no improvement is observed over a few epochs.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** |
| dyed-lifted-polyps | 0.8182 | 0.36 | 0.5 |
| dyed-resection-margins | 0.608 | 0.915 | 0.7305 |
| esophagitis | 0.5686 | 0.725 | 0.6374 |
| normal-cecum | 0.9296 | 0.33 | 0.4871 |
| normal-pylorus | 0.9401 | 0.785 | 0.8556 |
| normal-z-line | 0.6541 | 0.605 | 0.6286 |
| polyps | 0.4136 | 0.385 | 0.5637 |
| ulcerative-colitis | 0.7238 | 0.38 | 0.4984 |
| **Results** | | | |
| macro avg | 0.707 | 0.6231 | 0.6126 |
| weighted avg | 0.707 | 0.6231 | 0.6126 |
| accuracy | 0.6231 | | |

Table 2- Results Table for ViT with YOLO

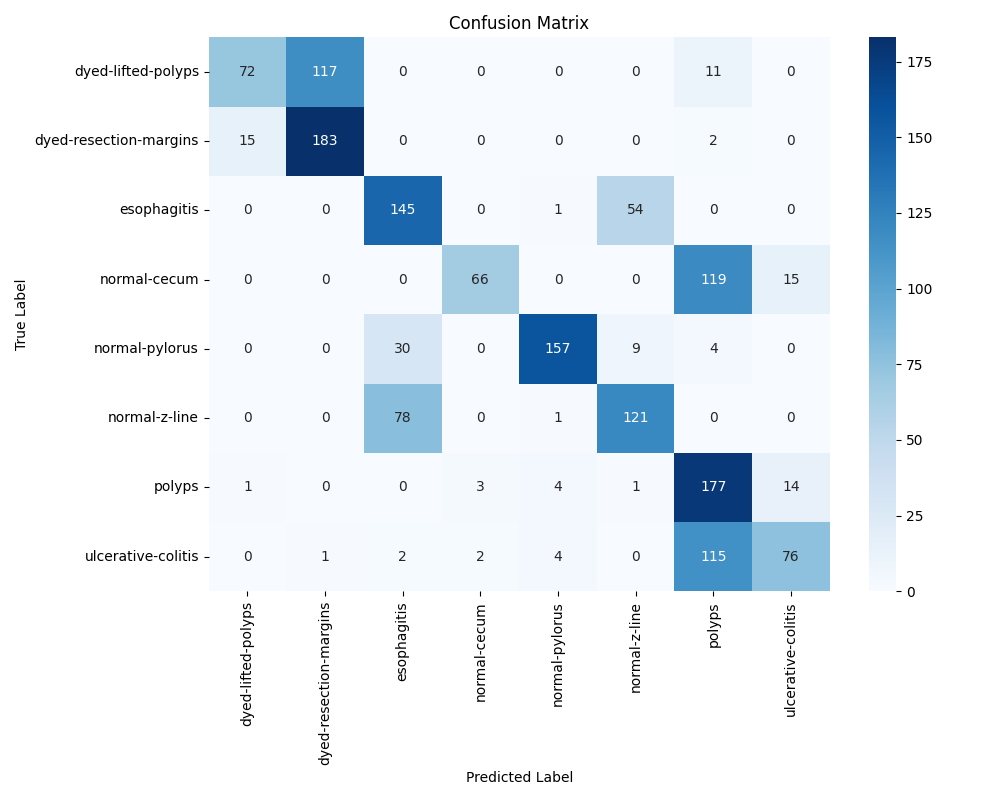


Figure 2- Confusion matrix for ViT with YOLO

Our implementation of a Vision Transformer model for classifying medical images from the Kvasir dataset demonstrated promising but mixed results across different diagnostic categories. Looking at Figure 2 and Table 2 we can observe several interesting patterns in the model's performance.

The model achieved an overall accuracy of 62.31%, which reflects the challenging nature of medical image classification. Breaking down the performance by category, we see notable variations in effectiveness. For instance, the model performed particularly well with "dyed-resection-margins," achieving a precision of 0.91 and an impressive recall of 0.92. This strong performance likely stems from the distinctive visual characteristics of dyed tissues that make these images more readily distinguishable.

However, some categories proved more challenging. The "polyps" category, while having high recall (0.89), showed lower precision (0.41), indicating a tendency to over-predict this condition. This is visible in the confusion matrix, where we can see substantial false positives, particularly with "normal-cecum" and "ulcerative-colitis" being misclassified as polyps.

Our pipeline's integration of ROI detection and data augmentation techniques appears to have helped in some cases but couldn't fully address all classification challenges. For example, the "esophagitis" category shows moderate performance with an F1-score of 0.64, suggesting that while the model can identify some characteristic features, there's room for improvement in distinguishing it from similar conditions.

The confusion matrix reveals interesting patterns of misclassification. There's notable confusion between related categories - for instance, "normal-z-line" and "esophagitis" show some mutual misclassification, which is understandable given their anatomical proximity and potential visual similarities. This suggests that while our ROI detection helps focus on relevant image regions, some subtle diagnostic differences remain challenging for the model to distinguish.

The macro-average metrics (precision: 0.71, recall: 0.62, F1-score: 0.61) indicate reasonably balanced performance across categories, though with clear room for improvement. These results suggest that while our approach successfully captures many relevant features through the ViT architecture and preprocessing pipeline, there might be benefit in exploring additional techniques to improve discrimination between visually similar conditions.

References

[1] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[2] P. Bissoonauth-Daiboo, M. Heenaye-Mamode Khan, M. M. Auzine, X. Gao, S. Baichoo, and Z. Heetun, “Endoscopic Image Classification using Vision Transformers,” in *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, Istanbul Turkiye: ACM, Oct. 2023, pp. 128–132. doi: 10.1145/3633598.3633623.

[3] J. Chen *et al.*, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” Feb. 08, 2021, *arXiv*: arXiv:2102.04306. doi: 10.48550/arXiv.2102.04306.

[4] M. Caron *et al.*, “Emerging Properties in Self-Supervised Vision Transformers,” May 24, 2021, *arXiv*: arXiv:2104.14294. doi: 10.48550/arXiv.2104.14294.

[5] M. C. H. Lee, O. Oktay, A. Schuh, M. Schaap, and B. Glocker, “Image-and-Spatial Transformer Networks for Structure-Guided Image Registration,” Jul. 22, 2019, *arXiv*: arXiv:1907.09200. doi: 10.48550/arXiv.1907.09200.

[6] S. Takahashi *et al.*, “Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review,” *J. Med. Syst.*, vol. 48, no. 1, p. 84, Sep. 2024, doi: 10.1007/s10916-024-02105-8.

[7] D. R. Sarvamangala and R. V. Kulkarni, “Convolutional neural networks in medical image understanding: a survey,” *Evol. Intell.*, vol. 15, no. 1, pp. 1–22, 2022, doi: 10.1007/s12065-020-00540-3.

[8] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[9] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial Transformer Networks,” Feb. 04, 2016, *arXiv*: arXiv:1506.02025. doi: 10.48550/arXiv.1506.02025.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” Mar. 18, 2021, *arXiv*: arXiv:2010.04159. doi: 10.48550/arXiv.2010.04159.

[12] Y. Dai, Y. Gao, and F. Liu, “TransMed: Transformers Advance Multi-Modal Medical Image Classification,” *Diagn. Basel Switz.*, vol. 11, no. 8, p. 1384, Jul. 2021, doi: 10.3390/diagnostics11081384.

[13] M. Assran *et al.*, “Masked Siamese Networks for Label-Efficient Learning,” Apr. 14, 2022, *arXiv*: arXiv:2204.07141. doi: 10.48550/arXiv.2204.07141.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” May 09, 2016, *arXiv*: arXiv:1506.02640. doi: 10.48550/arXiv.1506.02640.

[15] “Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets.” Accessed: Feb. 20, 2025. [Online]. Available: https://arxiv.org/html/2407.19394v4

[16] “OpenCV: Contours Hierarchy.” Accessed: Feb. 20, 2025. [Online]. Available: https://docs.opencv.org/4.x/d9/d8b/tutorial\_py\_contours\_hierarchy.html

[17] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, “Albumentations: fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.

Appendix A.  
  
An Example of an Appendix

Note: If you have figures or table in your appendices, do not include them in your List of Figures or List of Tables.

Number your figures and tables with the appendix letter e.g. Table A.1. (for the first table in Appendix A) or Figure B.2. (for the second figure in Appendix B).

Remember to delete this note before submission.