

Multi-Class Gastrointestinal Abnormality Detection using Transfer Learning with Swin Transformer

Manav Anand^a, Kshitij Singh^a

^a National Institute of Technology Delhi

Email: anandmanav1115@gmail.com, skshitij269@gmail.com

Abstract

In this study, Team Rookies presents a deep learning approach for the Capsule Vision 2024 Challenge, utilizing a fine-tuned Swin Transformer model for multi-class abnormality classification in video capsule endoscopy images. The dataset comprised over 50,000 frames from three public sources and one private dataset, labeled across ten gastrointestinal classes, including Angioectasia, Bleeding, Erosion, and more.

Our model achieved an overall accuracy of 93% on the validation set, with class-wise precision ranging from 0.65 to 0.99 and F1-scores between 0.50 and 0.98. Notably, classes such as Ulcer and Worms achieved F1-scores of 0.96 and 0.97, respectively, while Erythema and Erosion recorded lower F1-scores, indicating areas for improvement. These results underscore the potential of Swin Transformer-based architectures in enhancing the automated detection of gastrointestinal conditions, thereby facilitating early diagnosis and reducing manual review time in clinical practice.

As participants in the Capsule Vision 2024 Challenge, we fully complied with the competition's rules as outlined in the official guidelines. Our AI model development was based exclusively on the datasets provided in the official release. We are proud to have achieved 12th rank in the challenge, showcasing the effectiveness of our approach and its competitive performance.

Our code is available on [GitHub](#).

1 Introduction

Gastrointestinal (GI) and liver diseases are becoming increasingly common worldwide, driven by factors such as industrialization, changes in diet, and the overuse of antibiotics. These conditions present significant diagnostic and treatment challenges, underscoring the need for advanced medical technologies. Video Capsule Endoscopy (VCE) is a revolutionary non-invasive tool that enables the examination of the GI tract, particularly in diagnosing small intestine disorders like Crohn's disease, Celiac disease, and gastrointestinal cancer. Unlike traditional endoscopy, VCE employs a pill-sized camera that traverses the digestive system, capturing images and offering a detailed view of areas often inaccessible to conventional methods. This technique eliminates the need for sedation or invasive procedures, making it a preferred diagnostic tool for many patients.

Despite these advantages, VCE generates vast amounts of data—up to a million images per procedure over several hours—which gastroenterologists must manually review. This manual inspection, typically taking several hours, is labor-intensive and prone to human error due to visual obstructions like bubbles, debris, or food particles. Furthermore, the growing demand for these procedures compared to the number of available specialists results in delays in diagnosis and treatment.

Artificial intelligence (AI) has shown immense promise in automating the analysis of medical images, addressing these challenges by significantly reducing the time required for diagnosis while maintaining high accuracy. AI-based models have been successfully applied to various areas of medical imaging, but developing robust and generalizable models for multi-class classification of abnormalities in VCE images remains an active research area.

This paper proposes a deep-learning approach using a Swin Transformer model to classify gastrointestinal abnormalities from VCE frames. The Swin Transformer, known for its success in computer vision tasks, is a hierarchical vision transformer that processes images patch-based, making it well-suited for handling the large-scale, high-resolution nature of VCE data. The model efficiently captures local and global features by leveraging its self-attention mechanism across multiple scales, crucial for identifying subtle abnormalities within complex and varying GI environments.

The proposed method utilizes a pre-trained Swin Transformer fine-tuned on a dataset containing ten classes of GI conditions, including angioectasia, bleeding, erosion, erythema, and foreign bodies. We incorporate data augmentation techniques such as horizontal flips and image normalization to enhance model generalization. The model's performance is evaluated through standard metrics like accuracy, precision, recall, and F1-score, offering a comprehensive analysis of its classification capabilities. By automating the detection of abnormalities in VCE images, this study aims to alleviate the workload of gastroenterologists, improve diagnostic efficiency, and provide a reliable AI-assisted tool for clinical use.

2 Methods

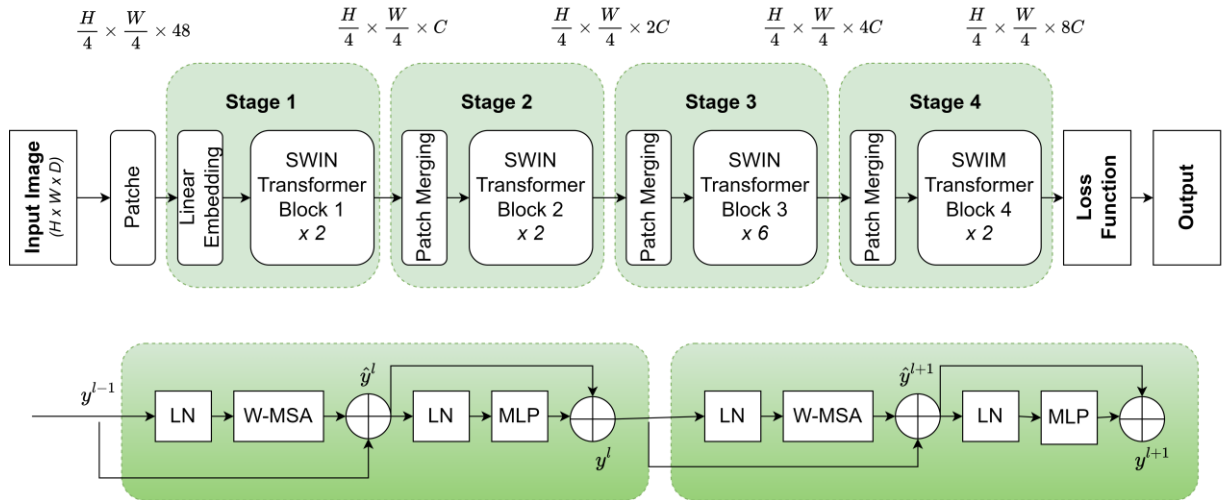


Figure 1: Block diagram of the proposed methodology for multi-class classification of gastrointestinal abnormalities in Video Capsule Endoscopy (VCE) frames.

This study employed a transfer learning-based approach using a pre-trained Swin Transformer model to classify images from a gastrointestinal abnormality dataset. The methodology involves several key stages: data preprocessing, model initialization, training, validation, and performance evaluation. Below is a detailed breakdown of the steps involved.

1. Model Architecture

We used a Swin Transformer model, specifically the Swin Tiny Patch4 Window7 224 architecture, pre-trained on the ImageNet dataset. The model was modified to accommodate our classification task by adjusting the output layer to classify ten distinct classes corresponding to various gastrointestinal abnormalities. The final layer of the Swin model was replaced with a fully connected layer with ten output nodes representing the ten classes.

2. Data Preprocessing

Data preprocessing was conducted using the `torchvision.transforms` module. The images in both the training and validation sets were resized to 224×224 pixels to match the input requirements of the Swin Transformer model. Additional augmentations were applied to the training data to improve generalization, including: **RandomHorizontalFlip**: This randomly flips the input images horizontally. **Normalization**: Each image was normalized using the mean and standard deviation of ImageNet-pretrained models ([0.485, 0.456, 0.406] for mean and [0.229, 0.224, 0.225] for standard deviation).

3. Dataset and DataLoader

The dataset used in this study consists of images from a Kaggle challenge dataset related to capsule endoscopy. The dataset was organized into training and validation folders and loaded using `torchvision.datasets.ImageFolder`. The `DataLoader` was employed to efficiently batch the data, with a batch size of 32, and the data was shuffled for the training set while keeping the validation set for evaluation.

4. Loss Function and Optimizer

For the classification task, `CrossEntropyLoss` was used as the loss function to handle the multi-class classification problem. The Adam optimizer was chosen for updating the model weights, with a learning rate set to 0.0001. The Adam optimizer helps achieve faster convergence during training due to its adaptive learning rate properties.

5. Training Procedure

The training consisted of feeding batches of images through the model and calculating the loss using the `CrossEntropyLoss` criterion. The gradients were backpropagated through the network, and the optimizer adjusted the model weights. Each epoch involved looping through the entire training set, and the model's performance was measured using accuracy. The training loop tracked both training loss and training accuracy after each epoch. To prevent overfitting, an early stopping mechanism was implemented, with a patience of 2 epochs. The training process was halted early if the validation loss did not improve for 2 consecutive epochs.

6. Validation Procedure

The model was evaluated on the validation dataset at the end of each epoch. The

validation procedure was similar to training but without gradient updates. The model's output was compared to the true labels to calculate the validation loss and accuracy. Additionally, all true and predicted labels were stored to generate a confusion matrix and a classification report at the end of training.

7. Handling Class Imbalance

To handle class imbalance in the gastrointestinal abnormality dataset, we implemented several mitigation techniques to ensure robust model training. First, we employed **class-weighted loss functions** using **CrossEntropyLoss**, which adjusted the importance of each class during training based on its frequency in the dataset. This approach helped the model focus more on underrepresented classes, ensuring fair learning across all categories. Additionally, we utilized **data augmentation techniques**, such as **random horizontal flips** and **image normalization**, to artificially increase the diversity of the training data, enhancing the model's ability to generalize to unseen variations. These strategies collectively helped mitigate the impact of class imbalance, improving the model's overall performance and ensuring it could accurately classify both common and rare gastrointestinal abnormalities.

3 Results

The experiments were conducted on a system running Microsoft Windows 11 Home, Version 10.0.22631, featuring an AMD Ryzen 7 4800H processor (2.90 GHz, 8 cores, 16 logical processors) and 16 GB of RAM. The BIOS version was FA506ICB.307, and the system had 15.4 GB of total physical memory and 29.9 GB of virtual memory. Kernel DMA protection and virtualization-based security were enabled for secure operations during model training.

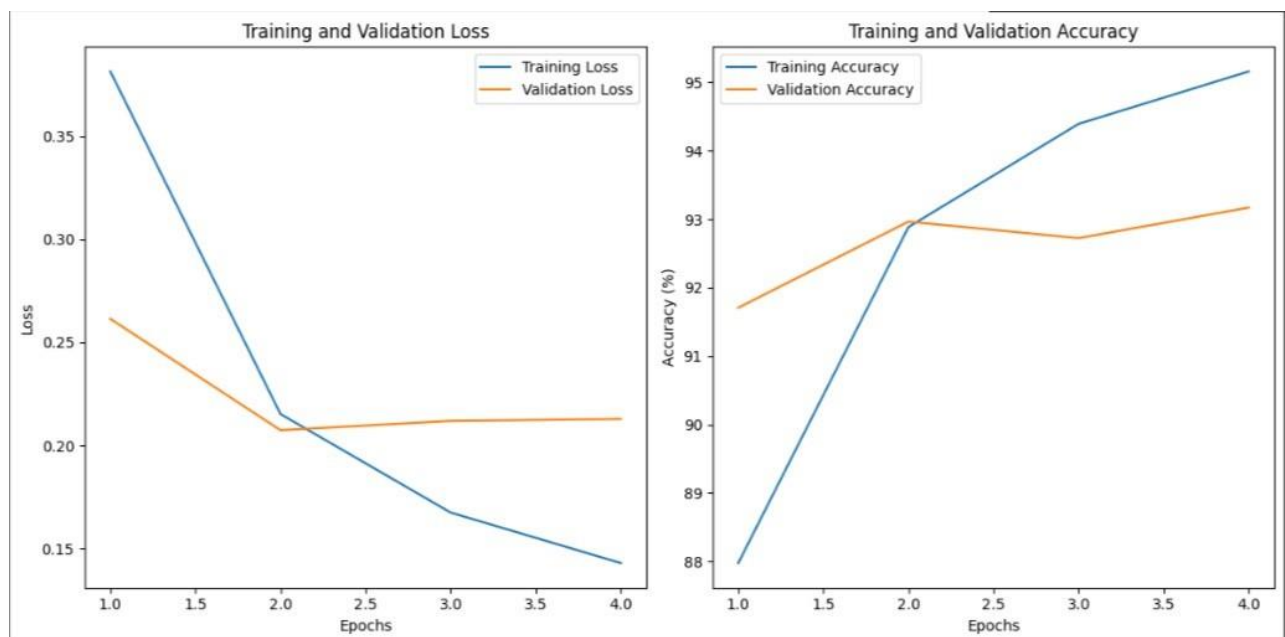
The primary programming language was Python, and PyTorch was used as the machine learning framework. The TimM library provided pre-trained models, and Matplotlib was used for data visualization.

During evaluation, training and validation metrics (loss and accuracy) were tracked to assess the performance of the Swin Transformer model. A confusion matrix summarized the classification outcomes. A detailed classification report provided precision, recall, F1 scores, and accuracy for all classes, highlighting the effectiveness of the model in classifying abnormalities in Video Capsule Endoscopy frames. The table ?? summarizes the hyperparameters used in the training of the model:

Table 1: Hyperparameters for Swin Transformer Model

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	0.0005
Weight Decay	0.05
Epochs	30
Batch Size	16
Image Size	224x224
Patch Size	4x4
Number of Layers	4
Number of Heads	3, 6, 12, 24
Dropout	0.1
Warmup Steps	500
Learning Rate Scheduler	Cosine Annealing
Gradient Clipping	1.0

Figure 2: The training and validation loss and accuracy when training dataset split into 80:20 ratio for 5 epochs



3.1 Achieved results on the validation dataset

The left plot illustrates the Training and Validation Loss over five epochs. The training loss consistently decreases, indicating that the model is effectively learning from the training data. In contrast, the validation loss also declines initially but levels off after the second epoch, suggesting that the model's ability to generalize to unseen data is not improving further. The right plot displays the Training and Validation Accuracy across the same epochs. The training accuracy rises sharply, reaching nearly 95% by the fourth epoch, demonstrating strong performance on the training set. However, the validation accuracy increases initially but then stabilizes around 93%, indicating a lack of further improvement in the model's performance on the validation set. Together, these plots suggest that while the model is learning well from the training data, it may be starting to overfit, as evidenced by the divergence between training and validation metrics after the second epoch.

The confusion matrix offers a detailed breakdown of correct and incorrect predictions for each class. Most values are concentrated along the diagonal, reflecting many correct classifications. For instance, the model classified normal findings with 97% accuracy and correctly identified ulcers with 98% accuracy. However, misclassifications occurred in more visually similar classes, such as between erosion and erythema, where the model occasionally needed clarification. This suggests that additional training data or more refined feature extraction techniques help the model distinguish between such challenging classes better. Overall, the confusion matrix shows strong performance across the majority of classes.

The classification report reveals a detailed performance evaluation of the Swin Transformer model on a dataset consisting of ten distinct classes related to gastrointestinal

Figure 3: The confusion matrix results on validation dataset

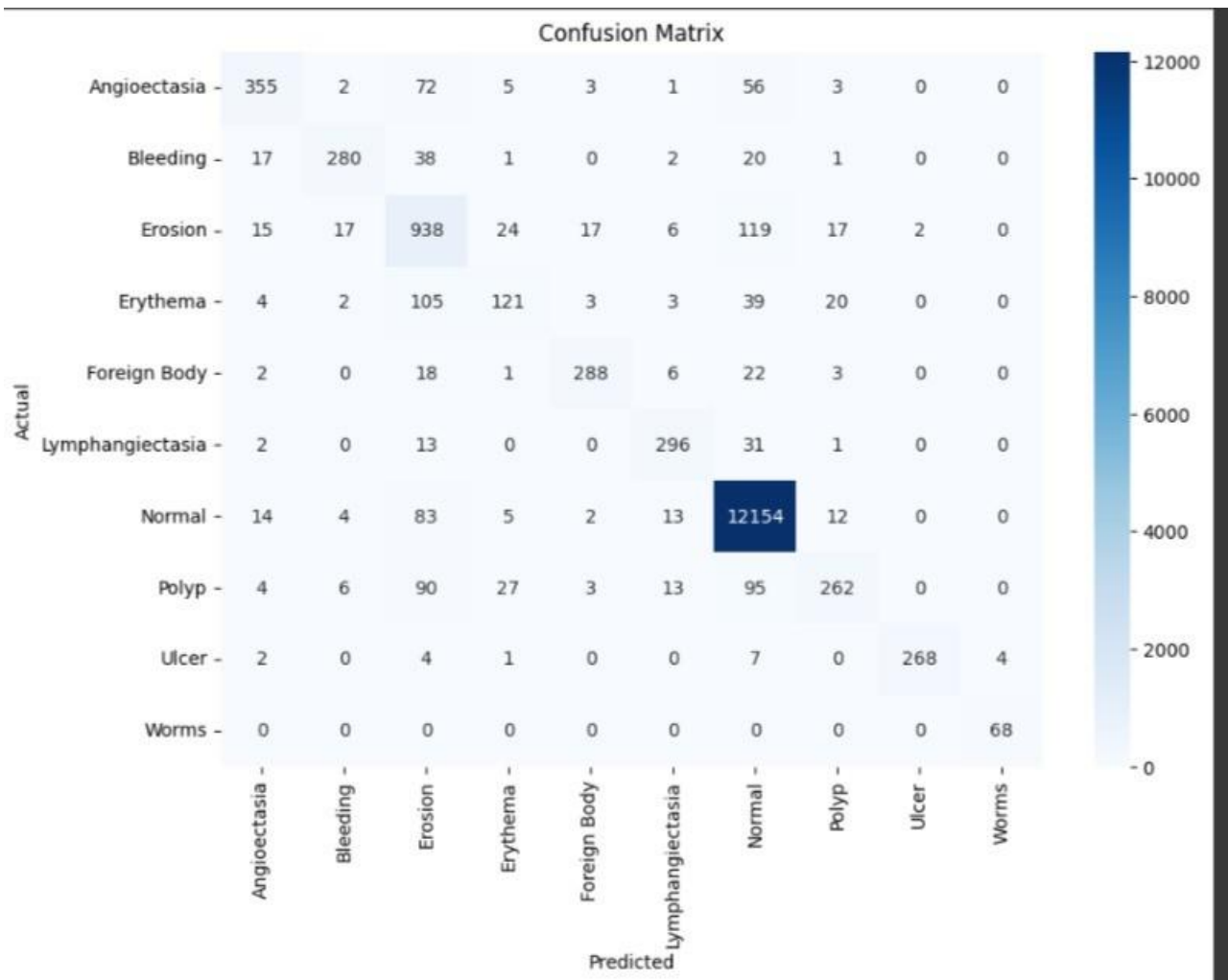


Figure 4: classwise precision, recall, f1-score and overall accuracy of the model on validation dataset

Classification Report:				
	precision	recall	f1-score	support
Angioectasia	0.86	0.71	0.78	497
Bleeding	0.90	0.78	0.84	359
Erosion	0.69	0.81	0.75	1155
Erythema	0.65	0.41	0.50	297
Foreign Body	0.91	0.85	0.88	340
Lymphangiectasia	0.87	0.86	0.87	343
Normal	0.97	0.99	0.98	12287
Polyp	0.82	0.52	0.64	500
Ulcer	0.99	0.94	0.96	286
Worms	0.94	1.00	0.97	68
accuracy			0.93	16132
macro avg	0.86	0.79	0.82	16132
weighted avg	0.93	0.93	0.93	16132

abnormalities. The model achieved an overall accuracy of 93%, indicating strong predictive performance.

Analyzing the precision, recall, and F1 scores for each class provides further insights:

1. Precision values range from 0.65 to 0.98, with the highest precision for the class "Ulcer" at 0.99. This means the model is particularly good at minimizing false positives for this class, effectively identifying true cases.
2. Recall scores vary significantly across classes, with the highest recall recorded for "Worms" at 1.00, indicating that the model successfully identifies all instances of this class. However, classes like "Erythema" have a lower recall of 0.41, suggesting some missed instances, which points to potential areas for improvement.
3. The F1 score, which balances precision and recall, further reflects the model's performance. The F1 scores range from 0.50 for "Erythema" to 0.97 for "Worms," indicating a solid balance between precision and recall for most classes but highlighting concerns for Erythema's performance.
4. The report also includes macro and weighted averages for precision, recall, and F1 score, offering a broader view of the model's performance across all classes. The macro average (considering all classes equally) and weighted average (taking into account class imbalance based on support) both suggest that the model performs well overall, with the weighted average F1 score at 0.93.

In summary, while the Swin Transformer demonstrates high accuracy and generally strong performance across most classes, the variability in recall and F1 scores across different classes signals the need for further optimization, particularly for classes where performance lags. This analysis can guide future refinements in the model's architecture or training strategies to enhance its effectiveness in detecting gastrointestinal abnormalities.

4 Discussion

In this research, we developed a deep learning framework utilizing the Swin Transformer architecture for the multi-class classification of gastrointestinal abnormalities in Video Capsule Endoscopy (VCE) images. The study addressed the growing need for efficient diagnostic tools in the context of increasing incidences of GI diseases, which often overwhelm healthcare systems due to the labor-intensive nature of VCE image analysis.

Our proposed model was trained on a well-structured dataset comprising various classes of GI abnormalities, including angioectasia, bleeding, erosion, erythema, and foreign bodies. The use of advanced data preprocessing techniques and augmentation strategies ensured that the model could generalize effectively across diverse visual conditions present in VCE images.

The performance evaluation revealed that the Swin Transformer achieved an overall accuracy of 93%, along with high precision and F1 scores for most classes. These results demonstrate the model's capability to significantly reduce the manual workload for gastroenterologists, enabling faster and more accurate diagnoses. However, certain classes, particularly "Erythema," exhibited lower recall and F1 scores, highlighting areas for future improvement.

Overall, this research underscores the potential of leveraging cutting-edge deep learning models like the Swin Transformer to enhance diagnostic accuracy in medical imaging. The findings pave the way for further exploration and refinement of AI-based tools, ultimately aiming to improve patient outcomes in gastrointestinal healthcare.

5 Conclusion

In conclusion, this research demonstrates the effectiveness of the Swin Transformer architecture in the multi-class classification of gastrointestinal abnormalities within Video Capsule Endoscopy (VCE) images. By harnessing advanced deep learning techniques, we successfully developed a robust model that achieved an accuracy of 93%, significantly enhancing the diagnostic capabilities for gastroenterologists.

The study addresses the pressing need for automated solutions to manage the increasing volume of VCE images, thereby reducing the burden on healthcare professionals while maintaining high diagnostic precision. The comprehensive evaluation of the model, including metrics such as precision, recall, and F1 scores, confirms its potential as a valuable tool in clinical settings.


6 Acknowledgments

As participants in the Capsule Vision 2024 Challenge, we fully comply with the competition's rules as outlined in [1]. Our AI model development is based exclusively on the datasets provided in the official release in [2]. We are proud to have achieved **12th rank** in the challenge, showcasing the effectiveness of our approach and its competitive performance.


References

- [1]  **Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, et al.** Capsule Vision 2024 Challenge: Multi-Class Abnormality Classification for Video Capsule Endoscopy. *arXiv preprint arXiv:2408.04940*, 2024.

URL: <https://arxiv.org/abs/2408.04940>.

- [2]  **Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Raman.** Training and Validation Dataset of Capsule Vision 2024 Challenge. *Figshare*, July 2024.

DOI: [10.6084/m9.figshare.26403469.v2](https://doi.org/10.6084/m9.figshare.26403469.v2).

- [3]  **Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Pallavi Sharma, Vijay Thakur, Dr. Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Ramanathan.** Testing Dataset of Capsule Vision 2024 Challenge. *Figshare*, October 2024.

DOI: [10.6084/m9.figshare.27200664.v3](https://doi.org/10.6084/m9.figshare.27200664.v3)