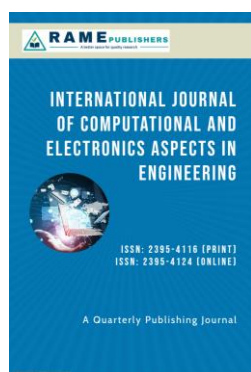


Structure-Aware Dual-View Contrastive Swin Transformer for Fine-Grained Gastrointestinal Endoscopic Image Classification

Fajer F. Fadhil* 

Network Unit, Computer Center, University of Mosul, Mosul, Iraq.

*Correspondence: fajrfehr@uomosul.edu.iq.



Abstract: The classification of gastrointestinal endoscopic images is still a difficult task due to the visual similarity of the related abnormalities and the high class imbalance. This study introduces a novel framework, SADC-Swin (Structure-Aware Dual-View Contrastive Swin), based on Swin Transformer V2, which incorporates dual-view supervised contrastive learning, ordinal supervision for ulcerative-colitis grades, and structure-aware supervision for normal Z-line and oesophagitis classes. The final number of images for experiments was 10,479 images from 16 classes after discarding classes with less than 100 images. Official dataset folds were kept and no exact duplicate images were found in or between folds using SHA-256 analysis. The conventional Swin-V2-T baseline model achieved 93.06% accuracy and 87.43% Macro-F1 on the validation fold, while SADC-Swin achieved 93.87% and 89.02%, respectively. On the independently reserved fold, the accuracy of SADC-Swin was 93.39% with a Macro-F1 score of 87.74% and the baseline was 92.82% and 86.11% respectively. The highest F1 improvements were seen in ulcerative-colitis grade 1 and oesophagitis-A. The results suggest that the clinically structured supervision could benefit the fine-grained endoscopic classification, especially for visually similar and under-represented classes.

Keywords: HyperKvasir, Swin Transformer, gastrointestinal endoscopy, contrastive learning, medical image classification.

Article – Peer Reviewed

Received: July 30, 2026

Accepted: September 16, 2026

Published: September 27, 2026

Copyright: © 2026 RAME Publishers

This is an open access article under the CC BY 4.0 International License.



<https://creativecommons.org/licenses/by/4.0/>

Cite this article: Fajer F. Fadhil, "Structure-Aware Dual-View Contrastive Swin Transformer for Fine-Grained Gastrointestinal Endoscopic Image Classification", *International Journal of Computational and Electronic Aspects in Engineering*, RAME Publishers, vol. 7, issue 4, pp. 479-497, 2026.

<https://doi.org/10.26706/ijceae.7.4.20260814>

1. Introduction

Gastrointestinal (GI) diseases encompass a broad spectrum of disorders of the gastrointestinal tract, and endoscopy is one of the most significant procedures for direct visualization of anatomical structures, inflammatory changes, polyps, and/or mucosal abnormalities and other pathological findings. But the ability to interpret the endoscopic images depends greatly on the experience of the endoscopist, and subtle visual differences may make the classification of similar abnormalities difficult. Over the past few years, artificial intelligence (AI), especially deep learning, has received growing attention as a means of supporting gastrointestinal endoscopy, such as for lesion detection, disease classification, severity assessment, and quality improvement. While there are significant challenges in generalizing these AI techniques, in gastrointestinal endoscopy, they have demonstrated remarkable promise, particularly in areas such as image analysis, diagnosis, and navigation. Clinical validation and interpretability are also key areas of concern that must be addressed before these tools can be widely adopted in clinical practice. [1], [2].

To develop reliable deep-learning systems for GI endoscopy, sufficient large and diverse image datasets are necessary. The HyperKvasir dataset is one of the largest publicly available gastrointestinal image analysis datasets. Borgli et al. presented HyperKvasir, which was created using 110,079 images and 374 videos from real gastro- and colonoscopy examinations. There are 23 classes in the labeled subset: 10,662 images are split into anatomical landmarks, pathological findings, normal findings, therapeutic interventions, and bowel-preparation quality. Some examples of the endoscopic classes in the present study are

shown in Figure 1, which demonstrates the high degree of visual diversity among the classes and the similarity between some of the clinically related classes. Official fold assignments are also included in the data set to enable a reproducible comparison of classification methods.[3],[4].

However, HyperKvasir is a difficult multiclass classification problem, due to the large differences in the number of images per class and the similarity in endoscopic features between some classes. Thus, overall accuracy alone may not be sufficient to capture the performance of a model, especially on more challenging or under-represented findings.

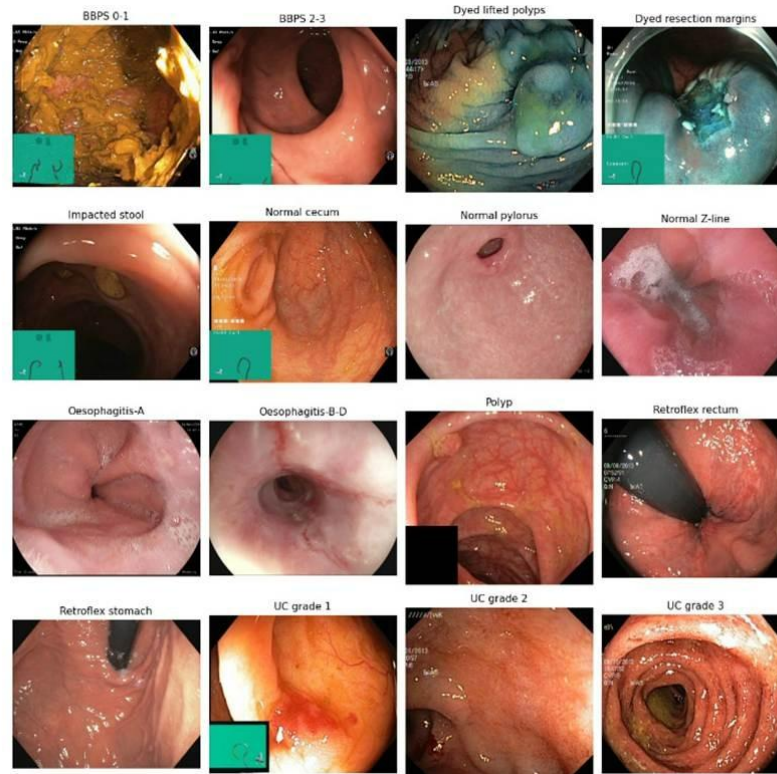


Figure 1. Representative endoscopic images from the 16 HyperKvasir classes considered in this study, including anatomical landmarks, pathological findings, bowel-preparation conditions, therapeutic appearances, and endoscopic viewing positions.

Convolutional neural networks (CNNs) have been the main players in medical image classification, due to their capacity to learn hierarchical visual representations directly from images. Self-attention mechanisms have also gained significant traction recently, as they enable modeling of relationships between spatially distant image regions more effectively, which is why Vision Transformers have been the focus of much research. The architectures based on transformers have therefore been explored for gastrointestinal disease classification, and recent works have shown the feasibility of such approach in the field of endoscopic image analysis [3]. Of these architectures, the Swin Transformer is a hierarchical architecture based on shifted-window self-attention, and the Swin Transformer V2 introduces improvements to the attention formulation, normalization, and position bias, which boosts training stability and scalability. [4],[5]. The properties of Swin-V2 make it a good choice for fine-grained endoscopic classification, where the diagnostic information may be present locally or over a wider area of the image.

However, a typical multiclassifier usually considers all the output classes as separate classes and tries to minimize the cross-entropy loss. This formulation does not explicitly take into account that some gastrointestinal labels have meaningful relationships. One good example is ulcerative colitis (UC) where the endoscopy may show varying degrees of severity. Deep-learning systems have been shown to help grade UC endoscopic activity in previous studies. Stidham et al. demonstrated that deep learning could perform performance similar to that of expert reviewers for endoscopic severity assessment, and subsequent multicenter and prospective studies also demonstrated the feasibility of automated prediction of the Mayo Endoscopic Subscore and the Ulcerative Colitis Endoscopic Index of Severity [6],[7],[8].

These studies also highlight the fact that the endoscopic severity grading might be challenging, as the visual features of adjacent stages can overlap, and that inter-observer variability is a significant problem.

The same problem is applicable to multiclass HyperKvasir classification. Classes 1, 2 and 3 ulcerative colitis are ordered classes, while the normal Z-line, oesophagitis-A and oesophagitis-B-D represent another group of visually related upper-GI classes. These relationships are not explicitly promoted by a conventional classification loss. Class weighting corrects for class frequency but will not directly enhance the separation of visually overlapping representations nor impose a structural relationship between disease categories. This indicates that the enhancement of feature discrimination and the introduction of available clinical label structure may be beneficial for hard multiclass endoscopy classification. Supervised contrastive learning is one possible way to enhance the quality of representation. In supervised contrastive learning, introduced by Khosla et al., the class of samples are grouped together in the embedding space, and the representations of different classes are separated. [9].

In contrast to traditional cross entropy learning, supervised contrastive learning is directly shaping the arrangement of the learned feature space. This is especially important for endoscopic images that have small visual variations among similar classes. Two independent augmented views of the same training image can also help to promote the learned representation to be consistent despite appearance changes and retain class-discriminative information.

In view of these factors, this study introduces a new multi-class gastrointestinal endoscopic image classification method, SADC-Swin (Structure-Aware Dual-View Contrastive Swin). The framework is based on Swin-V2-Tiny, the main backbone of the visual network, and adds three additional complementary mechanisms to the traditional supervised classification. First, the class representations are separated by applying dual-view supervised contrastive learning. Second, an auxiliary ordinal objective is added to the ulcerative-colitis grade classes to maintain the ordinal relationship between the classes when learning features. Third, for the related normal Z-line and oesophagitis categories, an auxiliary branch is used which is structure aware. The auxiliary targets are employed in training and the final inference is performed with the main multiclass classification head, without introducing unnecessary complexity in the prediction. The aim of the proposed framework is not just to augment the back-bone, but to test if the clinically structured supervision and contrastive representation learning can enhance the discrimination of challenging and visually similar gastrointestinal findings.

2. Literature Review

In recent years, the research of gastrointestinal endoscopic image analysis has moved from traditional CNN to attention-based CNN, Vision Transformer, self-supervised learning and hybrid feature-learning networks. While these approaches have yielded significant advances, multi-class classification is challenging due to the high visual similarity of related findings in the GI tract, the high-class imbalance and subtle differences in various stages of the same disease. These difficulties are especially evident in large datasets like HyperKvasir, which includes classes like normal anatomical landmarks, pathological conditions, therapeutic interventions, bowel preparation quality, and various grades of inflammatory disease.

Ramamurthy et al. [10] proposed Effimix, a multi-feature fusion architecture that integrates EfficientNet-B0 and a custom CNN with squeeze and excitation layers and self-normalizing layers. The model was tested on the HyperKvasir and reported to have an accuracy of 97.99%, precision of ~97% and recall of ~98%. The study showed that a fusion of complementary convolutional representations is effective in the classification for the GI endoscopy. The framework, however, is mostly CNN feature fusion and does not explicitly model relations between clinically related categories. In particular, inflammatory disease at various stages is treated as a separate label for the output, not as a part of an ordered or structurally related group of labels.

spantaleón-Pérez et al. [11] went into an empirical assessment of multiple attention-based architectures for automatic classification of gastrointestinal endoscopy in HyperKvasir. They explored recent attention and Transformer-based models instead of traditional CNNs. The best ensemble, which is the combination of two MobileViT models, achieved a Macro-F1 score of 0.636 with about 20 million parameters. Importantly, they found that the confusion-matrix analysis results indicated that the different ulcerative-colitis grades were not easily distinguished, even with the use of models that pay attention. This is very relevant as it suggests that the visual backbone is not sufficient to fully clear up the confusion among clinically related classes of disease severity. The study was mainly on architecture comparison and ensemble learning, and did not give any explicit ordinal supervision for the severity of ulcerative-colitis.

Wang et al. [12] proposed a Hybrid Shifted Windows based Vision Transformer for the classification of gastrointestinal endoscopy. They were inspired by the fact that the conventional Swin Transformer windows were not capable of capturing long-range relationships in complex endoscopic images. The proposed hybrid-window design was able to capture both short-range and long-range dependencies and reported accuracy of 95.42% on Kvasir v2 and 86.81% on HyperKvasir. Difficulties with widely distributed visual patterns (such as the normal Z-line and esophagitis) were specifically mentioned by the authors,

and were also seen in the present work. The hybrid shifted-window approach has been shown to be effective in spatial-context modeling, however, it is a modification of the attention mechanism itself, and not the explicit incorporation of clinical relationships between the normal Z-line and various categories of oesophagitis during training.

Ge et al. [13] explored deep-learning multiclassification of reflux esophagitis based on the Los Angeles classification. They used a set of images from HyperKvasir, as well as images from two hospitals, and classified them into normal, A+B for LA grades and C+D for LA grades. Multiple CNN and Transformer architectures, such as MobileNet, ResNet, Xception, EfficientNet, ViT, and ConvMixer, were compared. The most efficient results were reported by EfficientNet with validation and independent-test accuracy of 96.2% and 95.7% respectively, and with MCC value of 0.936 and 0.884 respectively. The study shows the potential of AI to accurately identify clinically relevant severity groups of oesophagitis. The range of its application was, however, limited to the classification of reflux-esophagitis and did not include the much wider multi-class GI problem in which oesophagitis had to be classified alongside of anatomical landmarks, polyps, ulcerative colitis, therapeutic findings and other endoscopic categories.

Stidham et al. [7] examined 16,514 images from 3,082 patients with ulcerative-colitis (UC) for automated grading of the severity of the disease based on the endoscopic findings. Their CNN-based model performed well with AUROC around 0.97 to differentiate between remission and moderate to severe disease, and it showed a high degree of agreement with the adjudicated Mayo scores, similar to that seen between experienced reviewers. For the present study, in particular, the authors considered Mayo severity to have ordinal properties instead of just being distinct categories. Their results show that there is useful information in the clinical ordering of disease severity labels for training a model. However, this task only used UC severity and did not include ordinal supervision as a part of a larger multi-class gastrointestinal classification task.

Guo et al. [14] proposed a curriculum self-supervised learning framework for HyperKvasir, which employed the vast amount of unlabeled data in the dataset. In their approach, they pretrained with ~100,000 unlabeled images, and used a modified labeled subset for downstream classification. To avoid extreme imbalance, 7 very small categories were eliminated, resulting in a 16-class classification setting that is very similar in scale to the present study. The proposed curriculum Mixup strategy outperforms vanilla SimSiam by 88.92% top-1 accuracy and 73.39% F1-score. This work showed the usefulness of representation learning for HyperKvasir, and revealed the challenge of the rare categories. Its contribution was to use self-supervised learning of a curriculum, rather than explicitly learning relationships between disease stages or related pathological categories during supervised fine-tuning.

Wang et al. [15] developed a foundation model, called Endo-FM, specifically for the analysis of endoscopy videos. The framework was based on a video Transformer trained using a large number of video clips (over 33,000) including up to 5 million frames, by self-supervised learning. Pretraining was done using global and local views to enhance robustness to spatial and temporal variations. Endo-FM outperformed the previous self-supervised and adapter-based methods in classification, segmentation and detection. This work suggests that representation learning for endoscopy is of significant value. The framework, however, requires video pre-training at a large scale, and it is mainly optimized for spatial-temporal analysis. The present problem is related to supervised classification of labeled still images, but differs in that it attempts to take advantage of the structure of the output labels without the need for large-scale external or unlabeled pretraining.

Comparative studies in recent years also show the complexity of the full-spectrum classification of HyperKvasir. In comparison, a 2026 comparison of ResNet50, EfficientNet-B0, ViT-B/16, and Swin Transformer-Tiny on all 23 labeled HyperKvasir classes, showed that the highest top-1 accuracy is obtained by Swin Transformer-Tiny (89.0%), but its Macro-F1 is only 0.587, which is much lower than the weighted F1 of 0.888. Class sizes with only a few samples did not fare well. The study found that class imbalance and performance for rare classes are still significant challenges even for the latest Transformer models. The experimental protocol varies slightly from the present study, but these results show the need for class-balanced metrics like Macro-F1 and balanced accuracy, and why overall accuracy alone is not sufficient for HyperKvasir. [16].

In general, the current literature shows some significant patterns. CNN feature-fusion methods can yield good classification performance; Transformer architectures can be leveraged to model contextual features; self-supervised approaches can make use of the large proportion of unlabeled data in HyperKvasir; and disease-specific systems can classify the severity of ulcerative-colitis or oesophagitis. Focusing on representation learning, however, relatively little attention has been paid to the integration of these ideas in a single multi-class framework, where the clinical structure of difficult labels is explicitly used. In particular, the confusion between ulcerative-colitis grades and between related upper-GI findings, namely normal Z-line and oesophagitis, have been reported repeatedly in previous HyperKvasir studies. These labels are treated independently

in standard cross-entropy classification and conventional class weighting focuses on the frequency instead of visual and clinical relationships.

Hence, in the current research, we propose a new SADC-Swin model, which is based on a robust Swin-V2-T visual backbone and dual-view supervised contrastive learning, ordinal supervision for ulcerative-colitis grades, and structure-aware supervision for the normal-Z-line/oesophagitis group. The proposed approach does not seek to expand the capacity of the model nor add another feature-fusion block, but rather attempt to guide the learned representation with respect to class identity and clinically relevant relationships between challenging classes. The final multiclass classifier is the only part that is needed at test time, and the auxiliary objectives are only required during training to enhance feature discrimination.

3. Material and Methods

The experimental setup in this study includes image preprocessing, a baseline model called Swin Transformer V2-Tiny (Swin-V2-T), and the proposed Structure-Aware Dual-View Contrastive Swin (SADC-Swin) architecture. The overall flow chart is shown in Figure 2. To emphasize that the difference in performance is mainly attributed to the proposed training strategy, rather than varying backbone capacity, both models share the same backbone, ImageNet-pretrained Swin-V2-T. SADC-Swin combines the dual-view supervised contrastive learning with two auxiliary objectives that are clinically relevant for gastrointestinal categories and are structure-aware. For inference, the principal 16-class classification head is only required.

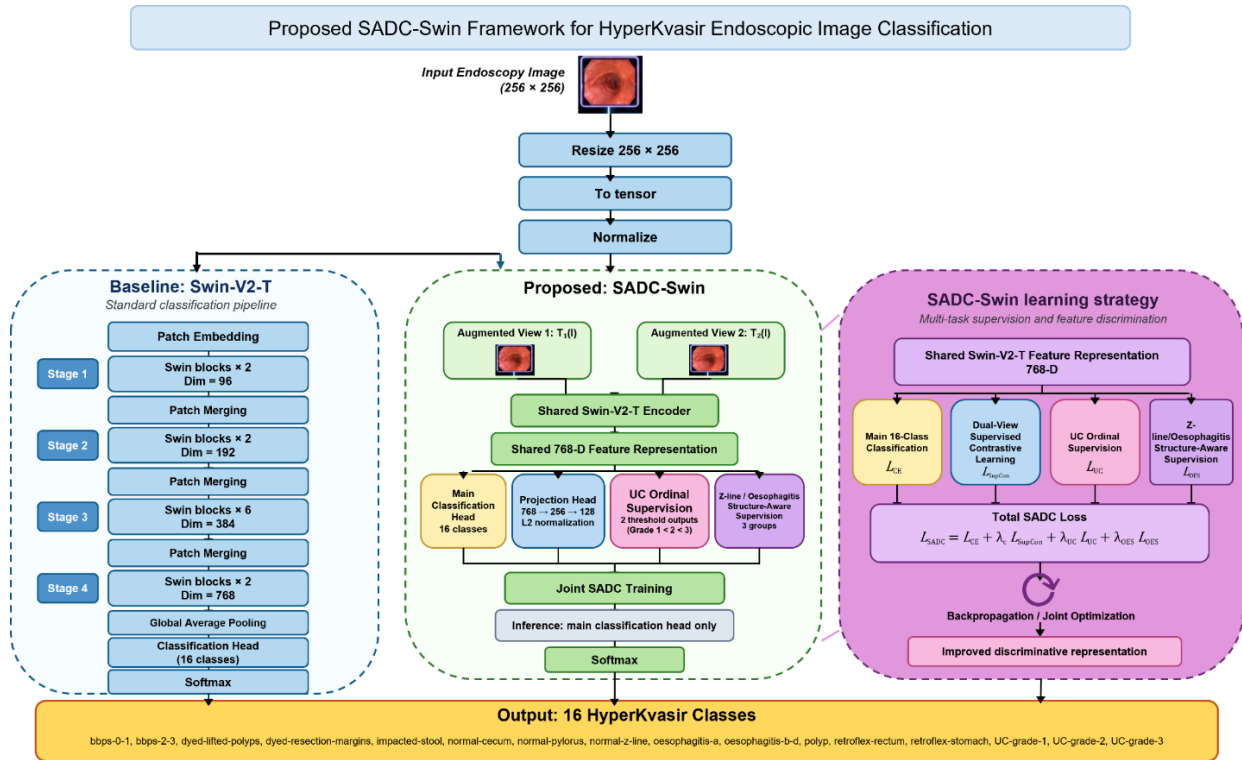


Figure 2. Overall workflow of the baseline Swin-V2-T and proposed SADC-Swin framework for 16-class gastrointestinal endoscopic image classification.

3.1 Dataset Description

The experiments were carried out on a publicly available gastrointestinal endoscopy dataset, HyperKvasir, which consists of 10,662 labeled images from 23 classes. Classes with fewer than 100 images were not included in the present classification study because of the very small number of images in several of the original categories. This gave an end result of 10,479 images in 16 classes.

The retained categories include various gastrointestinal observations such as anatomical landmarks, pathological observations, quality of bowel preparation, therapeutic appearances, and various endoscopic viewing positions. The distribution of the classes is given in Table 1.

Table 1. Distribution of the 16 HyperKvasir gastrointestinal endoscopic image classes included in this study.

Class	Number of Images
bbps-0-1	646
bbps-2-3	1,148
dyed-lifted-polyps	1,002
dyed-resection-margins	989
impacted-stool	131
normal-cecum	1,009
normal-pylorus	999
normal-z-line	932
oesophagitis-a	403
oesophagitis-b-d	260
polyp	1,028
retroflex-rectum	391
retroflex-stomach	764
ulcerative-colitis-grade-1	201
ulcerative-colitis-grade-2	443
ulcerative-colitis-grade-3	133
Total	10,479

There is a significant difference in the number of categories retained. For instance, there are 1,148 images in bbps-2-3, while impacted-stool has 131 images. Furthermore, certain classes are similar in clinical appearance, such as the three ulcerative-colitis grades, and the normal-Z-line/oesophagitis categories. Hence, overall accuracy was not deemed as a satisfactory measure for the evaluation of the model, and macro-averaged as well as class-specific measures were added.

3.2 Hardware and Software Configuration

Experiments were carried out in the Python programming language with the PyTorch deep-learning framework in a Linux-based JupyterLab environment. NVIDIA GB10 GPU with CUDA support was used for GPU acceleration. Table 2 provides a summary of the main software configuration.

Table 2. Hardware and software configuration used for model development and evaluation.

Component	Configuration
Deep-learning framework	PyTorch 2.10.0 + cu128
Torchvision	0.25.0 + cu128
CUDA	12.8
GPU	NVIDIA GB10
Development environment	JupyterLab
Programming language	Python 3.12
Mixed-precision training	FP16

Mixed-precision training was used to improve computational efficiency and reduce GPU-memory consumption during model optimization.

3.3 Image Preprocessing and Data Augmentation

All endoscopic images were converted to RGB format and transformed to an input resolution of 256×256 pixels, corresponding to the resolution used for the Swin-V2-T experiments.

The basic preprocessing pipeline can be represented as:

$$I \rightarrow \text{resize}(256 \times 256) \rightarrow \text{tensor} \rightarrow \text{normalization}$$

For training, moderate data augmentation was introduced to increase appearance variability and reduce overfitting. The augmentation procedure included random cropping, horizontal and vertical flipping, small-angle rotation, and mild color perturbation.

For the proposed SADC-Swin, each original training image I was independently augmented twice to produce two views:

$$I^{(1)} = T_1(I), \quad I^{(2)} = T_2(I)$$

where T_1 and T_2 denote independently sampled augmentation operations. The two views preserve the same class label while presenting slightly different visual appearances. These paired views are subsequently used by the supervised contrastive component of SADC-Swin. Validation images were processed using deterministic transformations without random augmentation.

3.4 Official Fold Assignment and Duplicate Analysis

The official five-fold split supplied with HyperKvasir was retained rather than generating a completely random image-level partition. During the development experiments, **fold 0** contained 2,088 images, **fold 1** contained 2,088 images, and the remaining three folds contained 6,303 images.

The development configuration was:

Training :2, 3, 4

Validation: fold 1

Reserved fold: fold 0

The distribution was therefore as shown in Table 3:

Table 3. Official-fold-based development partition of the 10,479 HyperKvasir images used in this study.

Partition	Images
Training	6,303
Validation	2,088
Reserved test fold (fold 0)	2,088
Total	10,479

In the model development phase, folds 2, 3, and 4 were used to train the model, and fold 1 was used to validate and select the model. Fold 0, which has 2,088 images, was not used for training, hyperparameter tuning, or selecting checkpoints. This reserved fold was only used once after the models were developed for the final independent evaluation of the selected Swin-V2-T baseline and SADC-Swin models.

An exact duplicate analysis was performed on the same machine before training, with the use of SHA-256 hashing. A cryptographic hash was computed for each image file and the same hashes were looked for in both the official folds and within each fold. Each of the 10,479 images generated a different hash, and there were no instances of exactly the same hash in the same fold or across folds. Therefore, there was no leakage found that was exactly the same in the selected images.

This integrity check is useful because if exact correlate and duplicate images are occurred across model-development partitions, it can result in an overly optimistic classification performance.

3.5 Baseline Model: Swin Transformer V2-Tiny

The main baseline was chosen to be Swin Transformer V2-Tiny. The weights were taken from the ImageNet-1K classification problem and then fine-tuned for the 16-class HyperKvasir classification problem.

Swin-V2 is a hierarchical Transformer network with four consecutive feature processing stages. The feature dimensions get larger roughly in the order of:

96 $\rightarrow$ 192 $\rightarrow$ 384 $\rightarrow$ 768. The corresponding numbers of Transformer blocks in the four stages are 2,2,6,2.

Following hierarchical feature extraction, the final representation is globally pooled and passed to a linear classification head 768 $\rightarrow$ 16.

For an input image I , the backbone produces a feature representation $h=E_{\theta}(I)$, $h \in \mathbb{R}^{768}$.

where E_{θ} represents the Swin-V2-T encoder.

The conventional classifier produces logits as in (1)

$$z=Wc+bc. \quad (1)$$

Where:

$$W_c \in \mathbb{R}^{16 \times 768}$$

The class probabilities are then obtained using Softmax as in (2):

$$P_c = \frac{\exp(z_c)}{\sum_{j=1}^{16} \exp(z_j)}. \quad (2)$$

The baseline model was optimized using standard multiclass cross-entropy loss as in (3):

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log(p_i, y_i). \quad (3)$$

This model serves as the reference architecture against which the proposed structure-aware training strategy is evaluated.

3.6 Proposed SADC-Swin Framework

The proposed SADC-Swin used Swin-V2 as a backbone with a modified supervision of the learned representation during training. The framework also adopts extra constraints to enhance discrimination between visually similar and clinically related endoscopic categories, apart from the sole use of a flat 16-class cross-entropy objective.

For each input image, the Swin-V2-T backbone generates a shared feature representation: $h=E_{\theta}(I)$, $h \in \mathbb{R}^{768}$

This common feature representation is provided to four training modules:

- the principal 16-class classification head;
- a supervised contrastive projection head;
- An ulcerative-colitis ordinal head;
- an oesophagitis/Z-line group head;
- The combined framework is trained together.

3.6.1 Main Multiclass Classification Head

The principal classification branch performs the original 16-class prediction as shown in (4):

$$Z_{\text{main}}=W_{\text{main}}h+b_{\text{main}} \quad (4)$$

The corresponding cross-entropy loss is denoted by $\mathcal{L}_{CE}$. This branch is retained during inference and therefore directly produces the final gastrointestinal class prediction.

3.6.2 Dual-View Supervised Contrastive Learning

To improve the discriminative organization of the learned representation space, two independently augmented views of each training image were passed through the same Swin encoder.

Let $I_i^{(1)}$, $I_i^{(2)}$ represent the two views of image i . Their backbone features are projected through a contrastive projection network $q=g(h)$, where the projection head consists of $768 \rightarrow 256 \rightarrow 128$.

GELU activation is used between the projection layers, and the resulting 128-dimensional representation is L2-normalized as in (5)

$$\tilde{q} = q / (\|q\|_2) \quad (5)$$

Supervised contrastive learning encourages representations belonging to the same class to become closer in feature space while pushing representations from different classes apart.

For anchor i , the supervised contrastive objective is expressed as in (6):

$$\mathcal{L}_{SupCon} = - \sum_{i=1}^N \frac{1}{|P(i)|} \sum_{p \in P(i)}^c \frac{\exp(\tilde{q}_i \cdot \tilde{q}_p / \tau)}{\sum_{a \neq i} \exp(\tilde{q}_i \cdot \tilde{q}_a / \tau)} \quad (6)$$

where $P(i)$ denotes the samples sharing the same class as anchor i , and the temperature was set to $\tau=0.10$.

This objective is intended particularly to improve the separation of endoscopic categories with overlapping visual characteristics.

3.6.3 Structure-Aware Ulcerative-Colitis Ordinal Supervision

The classes:

UC Grade1, UC Grade2, UC Grade3.

do not represent completely independent conditions but follow an ordered progression of endoscopic severity.

The three grades are encoded as: $Grade\ 1 = [0, 0]$, $Grade\ 2 = [1, 0]$, $Grade3 = [1, 1]$.

Accordingly, grade 2 requires the first severity threshold to be exceeded, whereas grade 3 requires both thresholds. Binary cross-entropy is applied to the ordinal outputs as in (7):

$$LUC = -1/NUC \sum [y \log \sigma(o) + (1-y) \log (1 - \sigma(o))]. \quad (7)$$

3.6.4 Oesophagitis/Z-Line Structure-Aware Supervision

A second clinically related group was identified among: normal-z-line, oesophagitis-a, and oesophagitis-b-d. These categories exhibit visual overlap around the gastroesophageal junction and can therefore be difficult to distinguish using only global multiclass supervision.

An auxiliary three-class head was therefore introduced as in (8):

$$O_{OES} = W_{OES} h + b_{OES}, \quad (8)$$

Where $W_{OES} \in \mathbb{R}^{3 \times 768}$

3.6.5 Joint SADC Objective

The final training objective integrates the four components as in (9):

$$\mathcal{L}_{SADC} = \mathcal{L}_{CE} + \lambda_c \mathcal{L}_{SupCon} + \lambda_{uc} \mathcal{L}_{UC} + \lambda_{OES} \mathcal{L}_{OES} \quad (9)$$

The coefficients used in the final configuration are $\lambda_c = 0.10$, $\lambda_{uc} = 0.20$, $\lambda_{OES} = 0.15$.

A key characteristic of SADC-Swin is that the auxiliary projection, ordinal, and oesophagitis heads are training-time components only. At inference, only Swin-V2-T backbone $\rightarrow$ main 16-class head is required. Thus, the proposed training strategy improves representation learning without requiring the auxiliary branches during final prediction.

3.7 Training Configuration

The proposed SADC-Swin and conventional baseline were trained with the AdamW optimizer. Differential learning rates were used to keep the pretrained visual representation and to enable faster learning of the newly introduced components.

The Swin-V2-T backbone used $LR_{backbone}=2 \times 10^{-5}$, whereas the newly introduced classification and auxiliary heads used $LR_{heads}=1 \times 10^{-4}$. Table 4 shows the final configuration for the training of SADC-Swin.

Table 4. Training configuration of the proposed SADC-Swin framework.

Parameter	Value
Optimizer	AdamW
Backbone learning rate	2×10^{-5}
New-head learning rate	1×10^{-4}
Weight decay	0.05
Scheduler	Cosine annealing
Epochs	15
Batch size	32 image pairs
Input resolution	256×256
Contrastive projection dimension	128
Supervised contrastive temperature	0.10
Contrastive loss weight	0.10
UC ordinal loss weight	0.20
Oesophagitis auxiliary loss weight	0.15
Contrastive warm-up	3 epochs
Gradient clipping	1.0
Mixed precision	FP16
Model-selection metric	Validation Macro-F1

A cosine-annealing learning-rate schedule was employed throughout the 15 epochs, with the minimum learning rate set to 1×10^{-6} .

The model checkpoint providing the highest validation Macro-F1 was retained as the final development-stage model.

4. Evaluation Metrics

Since the HyperKvasir classification problem used in this study has an imbalanced class distribution with 16 classes, the performance of the models was assessed using various complementary metrics, not only accuracy. The accuracy, precision, recall, specificity, class-wise F1-score, Macro-F1, weighted F1-score, balanced accuracy and Matthews correlation coefficient (MCC) were used in the evaluation. When calculating the class-specific measures the multiclass problem was addressed with a one-versus-rest approach.

Let $C=16$ denote the number of classes. For each class, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the proportion of correctly classified endoscopic images among all evaluated images [17],[18] as in (10).

$$Accuracy = \frac{\sum_{c=1}^C TP_c}{N} \quad (10)$$

While accuracy is a good overall measure of classification performance, it may be affected by the number of samples in each class. Hence, other class-balanced metrics were added as well.

Precision is the percentage of images correctly classified as a specific class as in (11):

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (11)$$

High precision means that the classifier makes few false positive classifications for a class.

Sensitivity (also known as recall) is the fraction of true positives that are detected as in (12):

$$Recall_c = \frac{TP_c}{TP_c + FN_c}. \quad (12)$$

Recall is of special significance for the less represented gastrointestinal categories as a high overall accuracy can be accompanied by poor recognition of minority classes [19],[20],[21].

Specificity is the accuracy of the classifier in rejecting images that are not in a specific class as in (13):

$$Specificity_c = \frac{TN_c}{TN_c + FP_c}. \quad (13)$$

For the multiclass problem, specificity was determined with a one-versus-rest interpretation for each class.

The F1-score represents the harmonic mean of precision and recall as in (14):

$$F1_c = \frac{2 Precision_c Recall_c}{Precision_c + Recall_c}. \quad (14)$$

The class-specific F1-score was used to examine performance on difficult categories such as ulcerative-colitis grades and oesophagitis classes [22],[23],[24].

Macro-F1 gives every class the same importance regardless of the number of samples belonging to that class. It is calculated as in (15):

$$Macro - F1 = \left(\frac{1}{C}\right) \sum_{c=1}^C F1_c. \quad (15)$$

Macro-F1 was selected as the principal model-selection metric in this study because the HyperKvasir subset is considerably imbalanced. Therefore, an improvement in Macro-F1 reflects improved classification across the class set rather than performance dominated by the largest categories.

Weighted F1 takes the number of samples in each class into account as (16):

$$Weighted - F1 = \sum_{c=1}^C \left(\frac{N_c}{N}\right) F1_c. \quad (16)$$

4.1 Statistical Analysis

Two complementary analyses were conducted to determine if the differences in performance between the conventional Swin-V2-T baseline and SADC-Swin were supported statistically on the independent reserved fold. First, an exact McNemar test was used on the two paired correct and incorrect predictions of the two classifiers. Secondly, a nonparametric bootstrap procedure, based upon 5,000 resamples, was used to estimate the 95% confidence interval for the difference in Macro-F1 between SADC-Swin and the baseline. For reproducible results a fixed random seed of 42 was used. The level of 0.05 was used to determine statistical significance.

5. Results and Discussion

5.1 Overall Classification Performance

The proposed SADC-Swin was compared with the conventional Swin-V2-T baseline, using the same training and validation partitions. For the principal model-selection measure, macro-F1 was used since there is a wide range of class frequencies. Table 5 shows the overall validation performance of both models.

Table 5. The overall validation performance of both models.

Metric	Swin-V2-T Baseline	Proposed SADC-Swin	Improvement
Accuracy	93.06%	93.87%	+0.81 pp
Macro-F1	87.43%	89.02%	+1.59 pp
Weighted F1	92.75%	93.76%	+1.01 pp
Balanced Accuracy	87.78%	89.30%	+1.52 pp
MCC	0.9246	0.9334	+0.0088

In general, SADC-Swin achieved better performance in all five overall evaluation metrics than the conventional Swin-V2-T baseline. The accuracy went up from 93.06% to 93.87% and the Macro-F1 score from 87.43% to 89.02% that is 1.59 percentage points.

The balanced accuracy also improved from 87.78% to 89.30%, and weighted F1 improved from 92.75% to 93.76%. The MCC also increased from 0.9246 to 0.9334, which means that the overall agreement between predicted and true class labels was better. The differences between the principal metrics are shown in Figure 3.

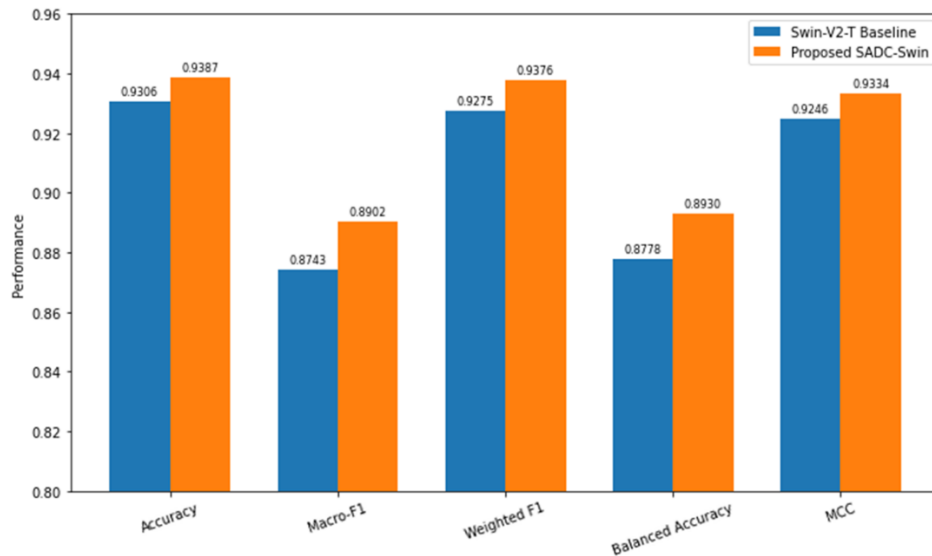


Figure 3. Overall validation performance comparison between the conventional Swin-V2-T baseline and the proposed SADC-Swin across accuracy, Macro-F1, weighted F1, balanced accuracy, and MCC.

5.2 Class-Wise Performance

The effect of SADC-Swin was also assessed individually for each of the 16 gastrointestinal categories. The F1-scores of the baseline and proposed models in each class are shown in Table 6.

Table 6. Class-wise F1-score comparison between Swin-V2-T and SADC-Swin

Classes	Baseline F1	SADC-Swin F1	Change
Ulcerative-colitis-grade-1	52.46%	62.16%	+9.70 pp
Oesophagitis-a	57.35%	66.67%	+9.31 pp
Ulcerative-colitis-grade-3	76.92%	79.25%	+2.32 pp
Dyed-lifted-polyps	94.76%	96.74%	+1.98 pp
Polyp	95.77%	97.58%	+1.81 pp
Dyed-resection-margins	95.14%	96.71%	+1.57 pp
Normal-z-line	89.62%	90.82%	+1.20 pp
bbps-0-1	98.44%	98.82%	+0.39 pp
Ulcerative-colitis-grade-2	81.72%	81.82%	+0.10 pp
Oesophagitis-b-d	81.90%	82.00%	+0.10 pp
Retroflex-rectum	96.73%	96.77%	+0.04 pp
bbps-2-3	97.54%	97.55%	+0.01 pp
Normal-cecum	99.00%	98.51%	-0.49 pp
Impacted-stool	82.54%	81.97%	-0.57 pp
Normal-pylorus	99.25%	98.26%	-0.99 pp
Retroflex-stomach	99.67%	98.68%	-0.99 pp

The proposed model achieved better F1 score in 12 classes out of 16. The largest improvement was seen with ulcerative-colitis-grade-1 with a 9.70 percentage-point increase from 52.46% to 62.16% for F1. The large improvement was achieved for oesophagitis-a with an F1-score rising from 57.35% to 66.67% (9.31 percentage points).

The findings are especially relevant since these two classes were among the hardest for the baseline, conventional Swin-V2. The same was achieved for ulcerative-colitis-grade-3, normal-z-line, polyps, and the two therapeutic categories, dyed. Small decreases were seen in normal-cecum, impacted-stool, normal-pylorus and retroflex-stomach. But the drops were only about 1 percentage point or less. Furthermore, the F1-scores of several of these categories were already near 99% for the baseline, indicating that there was comparatively little potential for improvement.

The overall class-wise comparison is shown in Figure 4.

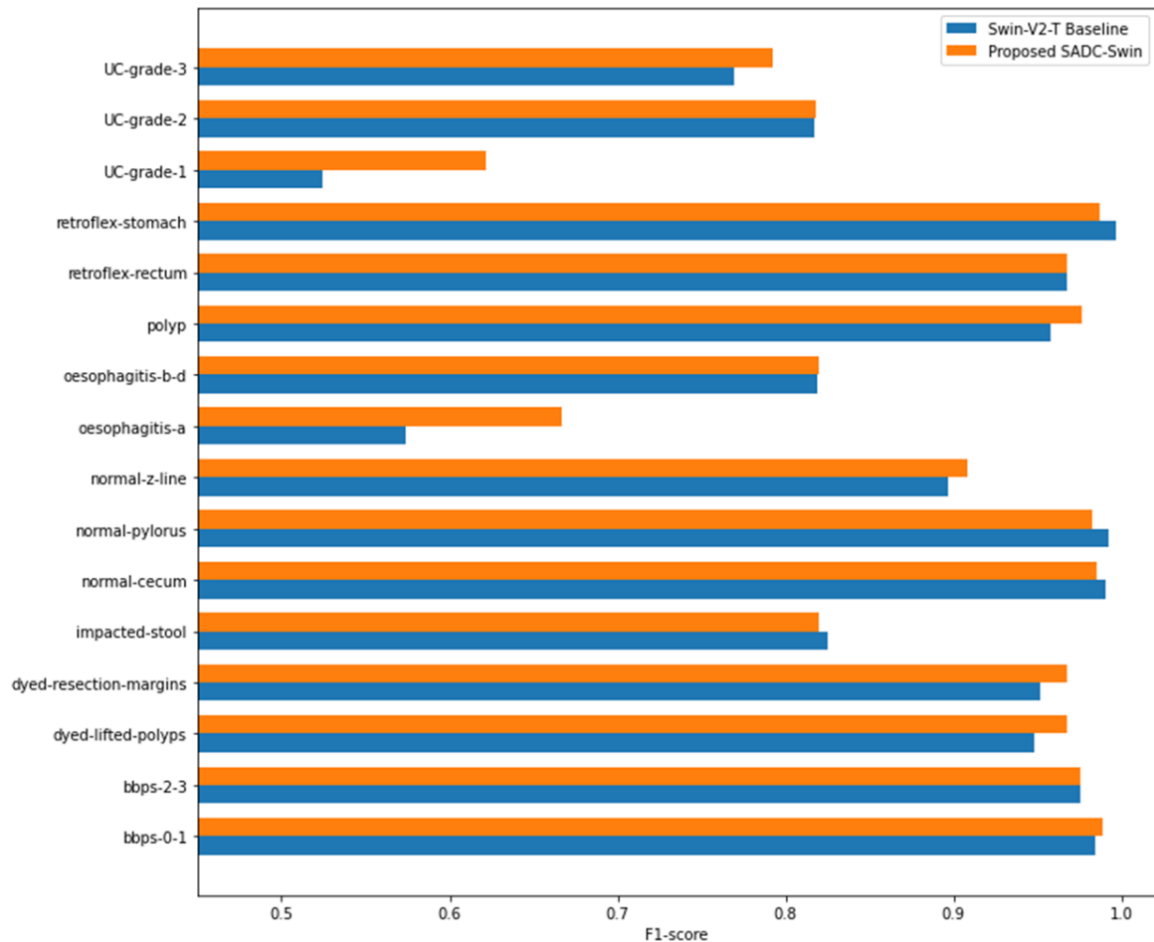


Figure 4. Class-wise validation F1-score comparison between the Swin-V2-T baseline and proposed SADC-Swin across the 16 HyperKvasir categories.

5.3 Improvement in the Most Challenging Classes

The most significant gains were seen in classes where patterns were clinically related or visually similar to each other, with SADC-Swin providing the greatest gains in these classes. This is seen more in ulcerative colitis and oesophagitis. For ulcerative-colitis-grade-1, F1:52.46%→62.16%. While oesophagitis-A improved from: 57.35%→66.67%.

The improvement in normal-Z-line from 89.62% to 90.82% is also remarkable as this class is part of the upper GI group targeted by the structure-aware auxiliary objective which is visually related. SADC-Swin showed an improvement in grade-1 and grade-3 performance for the ulcerative-colitis group, and essentially maintained the grade-2 performance. This behaviour facilitates ordinal supervision as the model is encouraged to learn these labels as severity-related classes instead of unrelated classes. The largest positive class-wise changes are highlighted in Figure 5.

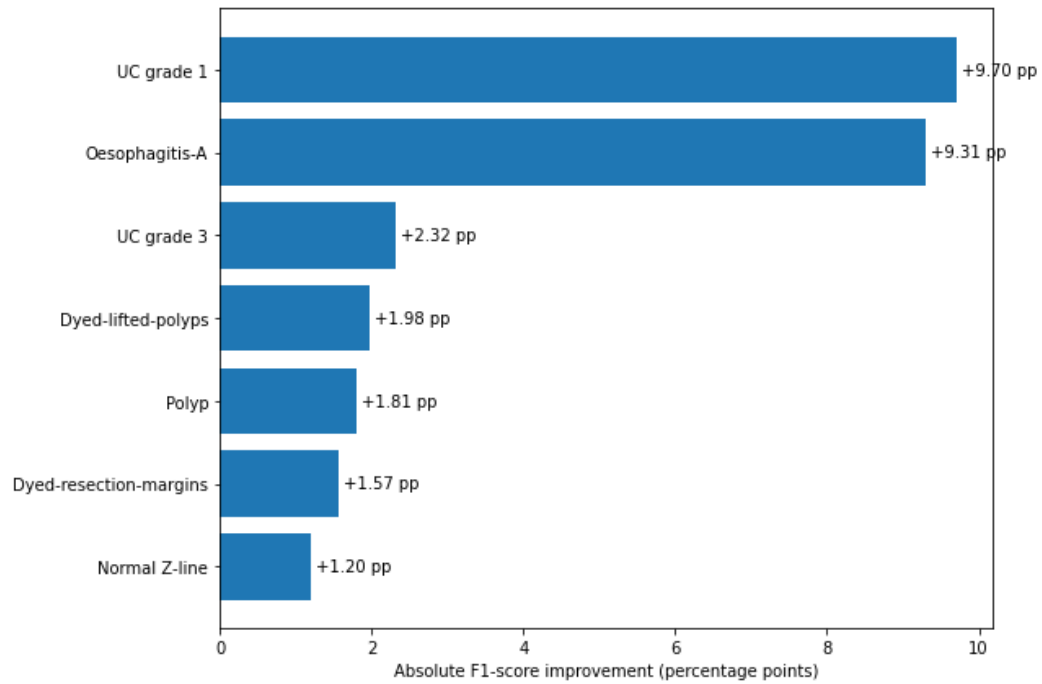


Figure 5. Largest class-specific F1-score improvements achieved by SADC-Swin relative to the Swin-V2-T baseline on the validation set.

5.4 Comparison of Training Strategies

A series of Swin-V2-T training strategies were tested to see if the performance improvement could be obtained through simpler modifications of the baseline. The validation Macro-F1 value of the conventional Swin-V2-T with cross-entropy was 0.8743, and the weighted cross-entropy slightly increased to 0.8744, as shown in Figure 6. DAPC-Swin boosted the Macro-F1 to 0.8781 and Prototype-Swin further raised it to 0.8816. The proposed SADC-Swin obtained the best validation Macro-F1 value of 0.8902, showing that the dual-view contrastive learning with structure-aware auxiliary supervision was the most effective training strategy among all the investigated training strategies.

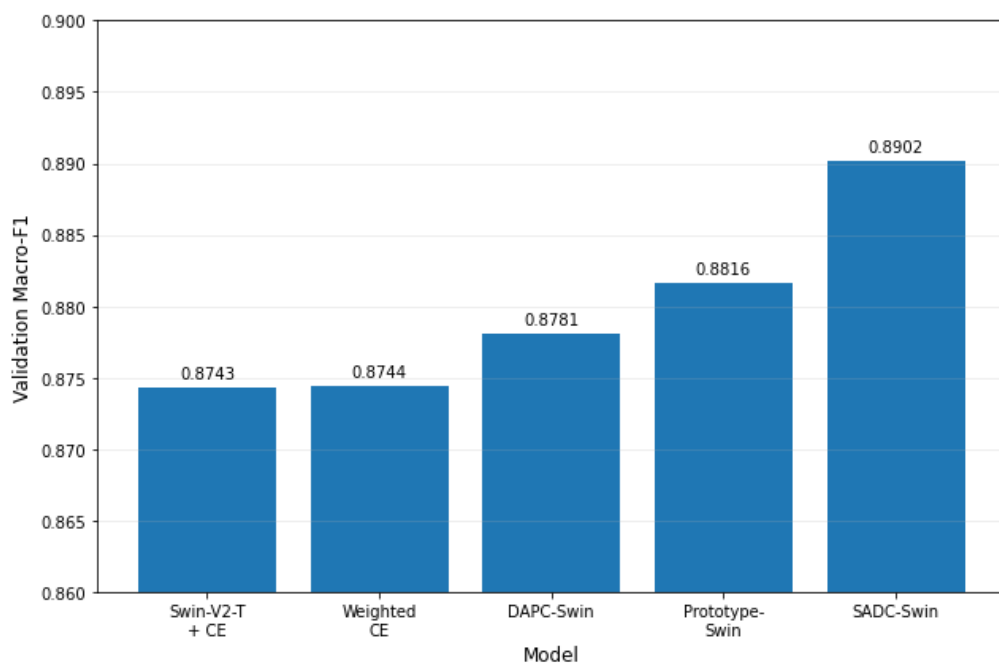


Figure 6. Validation Macro-F1 comparison of the investigated Swin-V2-T training strategies. The proposed SADC-Swin achieved the highest Macro-F1 of 0.8902, compared with 0.8743 for the conventional cross-entropy baseline.

Table 7 summarizes these development-stage experiments.

Table 7. Comparison of alternative training strategies investigated during model development

Model	Accuracy	Macro-F1	Balanced Accuracy	MCC
Swin-V2-T + CE	0.9306	0.8743	0.8778	0.9246
Swin-V2-T + weighted CE	0.9296	0.8744	0.8914	0.9235
DAPC-Swin	0.9325	0.8781	0.8772	0.9266
Prototype-Regularized Swin	0.9330	0.8816	0.8859	0.9271
Proposed SADC-Swin	0.9387	0.8902	0.8930	0.9334

5.5 Confusion-Matrix Analysis

The normalized confusion matrices for the baseline Swin-V2 and proposed Swin SADC are presented in Figure 7 and Figure 8, respectively. The clinically related oesophagitis / Z-line and ulcerative-colitis groups were the most prone to the main errors of the baseline. The correct classification rate rose from 49% to 60% and the confusion with normal Z-line fell from 38% to 33% and confusion with oesophagitis-B-D fell from 12% to 7% for oesophagitis-A.

The highest gains were seen in ulcerative-colitis grade 1, where the rate of correctly classified cases rose from 40% to 57%, and the rate of incorrectly classified cases dropped from 35% to 20% in grade 2. The classification rate in UC grade 3 also increased from 77% to 81%. There was still some residual confusion for UC grade 2 and oesophagitis-B-D, suggesting that adjacent clinical categories still had overlapping visual characteristics.

The overall confusion matrices show that the primary improvement in discrimination was in the difficult class groups that SADC-Swin was designed to tackle, which is consistent with the observed improvement in Macro-F1 and balanced accuracy.

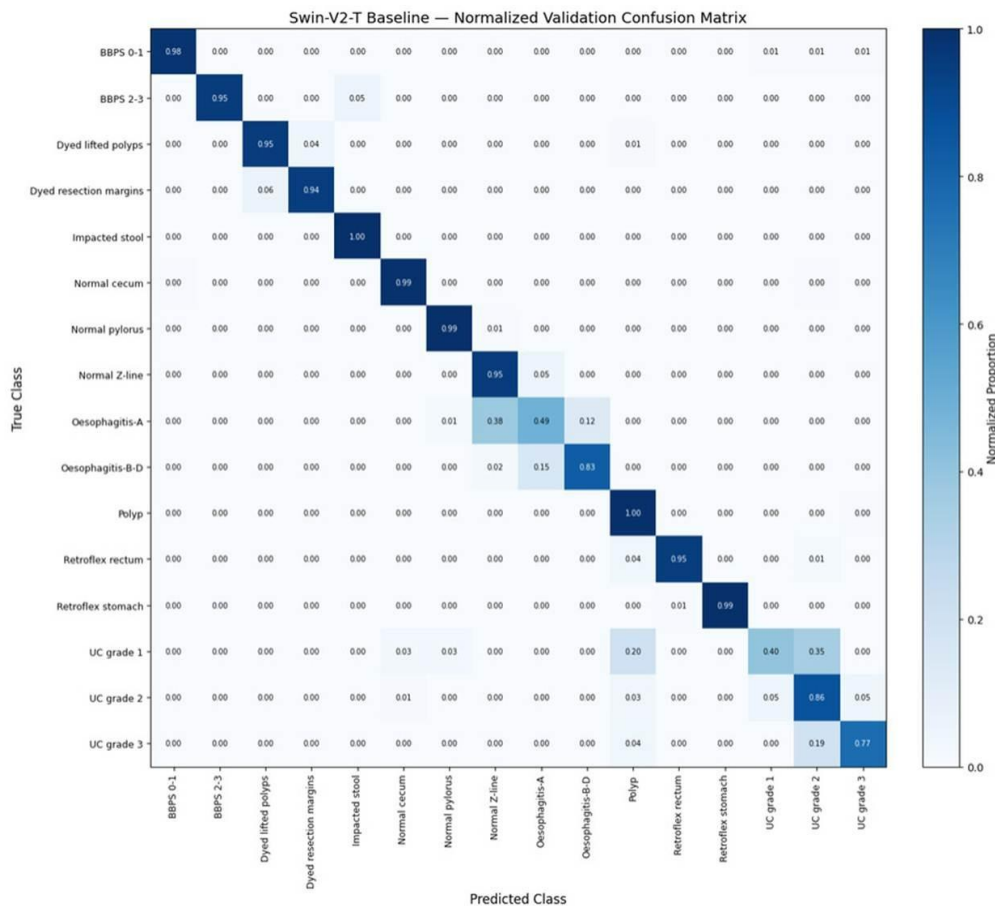


Figure 7. Normalized validation confusion matrix of the conventional Swin-V2-T baseline across the 16 gastrointestinal endoscopic classes.

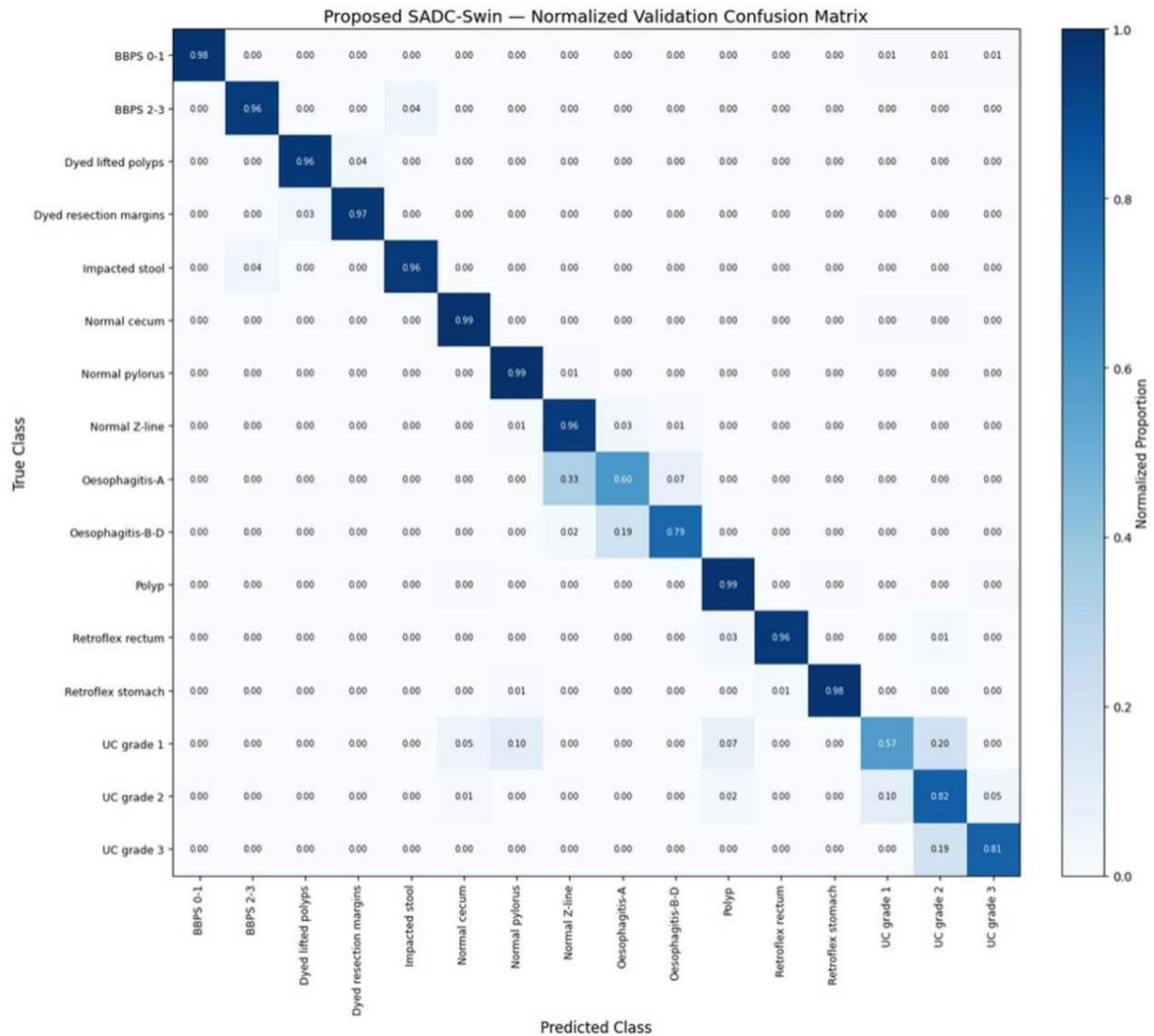


Figure 8. Normalized validation confusion matrix of the proposed SADC-Swin across the 16 gastrointestinal endoscopic classes.

5.6 Independent Reserved-Fold Evaluation

To determine if the improvement gained in model development extended to a fold that had not been used for validation in the process of selecting the checkpoints, the two models, Swin-V2-T baseline and SADC-Swin, were tested on the independently reserved official fold 0. This fold had 2088 images, which were not used for training, hyperparameter tuning, or model selection.

The conventional Swin-V2-T baseline was able to reach an accuracy of 92.82%, Macro-F1 of 86.11%, weighted F1 of 92.56%, balanced accuracy of 85.78%, and MCC of 0.9219 on the reserved fold. The proposed SADC-Swin achieved an accuracy of 93.39%, Macro-F1 of 87.74%, weighted F1 of 93.31%, balanced accuracy of 87.27%, and MCC of 0.9281 as shown in Table 8.

SADC-Swin thus achieved an improvement in accuracy of 0.57 percentage points, Macro-F1 of 1.63 percentage points, weighted F1 of 0.75 percentage points, and balanced accuracy of 1.49 percentage points, while MCC improved by 0.0062. Notably, the 1.63-percentage-point Macro-F1 improvement on the independent fold was similar to the 1.59-percentage-point improvement observed on the development validation fold.

The exact McNemar test showed 24 images that were correctly classified by the baseline only and 36 that were correctly classified by SADC-Swin only. The two-sided test gave P value of 0.155, showing that the difference in the accuracy of the paired classification was not statistically significant at the 0.05 level. The 95% CI for the Macro-F1 difference, however, was 0.13 to 3.15 percentage points with a bootstrap analysis. The improvement in Macro-F1 was supported by the bootstrap analysis as this interval did not include zero.

Table 8. Independent performance comparison on the reserved official fold 0.

Metric	Swin-V2-T Baseline	Proposed SADC-Swin	Difference
Accuracy	92.82%	93.39%	+0.57 pp
Macro-F1	86.11%	87.74%	+1.63 pp
Weighted F1	92.56%	93.31%	+0.75 pp
Balanced Accuracy	85.78%	87.27%	+1.49 pp
MCC	0.9219	0.9281	+0.0062

5.7 Discussion

The results of the experiments suggest that the main benefit of SADC-Swin is that it enhances the performance of the difficult classes and does not significantly affect the performance of the easy classes. The conventional Swin-V2-T baseline already had several anatomical and procedural classes that had F1-scores >95%. It was not therefore expected, nor necessary, that there would be any significant improvement in these classes.

The best benefit, however, was seen in the ulcerative-colitis-grade-1 and oesophagitis-A classes, the weakest of the standard classifier. The differences in their performance on F1 tasks, 9.70% and 9.31% respectively, are quite high compared to the differences in the easier tasks. This is consistent to the motivation of SADC-Swin, which is to alter the supervision of the shared feature representation during training, without increasing the size of the model.

This is further confirmed by the increase in both Macro-F1 and balanced accuracy. Macro-F1 improved by 1.59 percentage points, and balanced accuracy improved by 1.52 percentage points, suggesting that the improvement was not just due to the majority classes. Concurrently, the conventional accuracy rose by 0.81 percentage points, showing that the discrimination between the minority and difficult classes did not compromise the overall classification results.

The independent reserved-fold evaluation further corroborates the generalization of the proposed training strategy. SADC-Swin showed improvement on all five overall metrics with a 1.63 percentage point gain in Macro-F1 and a 1.49 percentage point gain in balanced accuracy. The difference in accuracy between the two was not statistically significant based on the exact McNemar test ($p=0.155$), but the 95% bootstrap confidence interval for the Macro-F1 improvement was still positive (0.13–3.15 percentage points). The results suggest that the main advantage of SADC-Swin is more apparent in terms of class-balanced performance than in terms of overall accuracy.

In conclusion, the results indicate that dual-view supervised contrastive learning with structure-aware auxiliary supervision can yield more discriminative representations for fine-grained gastrointestinal endoscopic image classification. The proposed auxiliary branches are only used during training, and the inference is still performed by the normal 16-class classification head. Hence, in deployment there is no need for a more complex multi-branch decision process.

6. Limitations

This study has several drawbacks that must be taken into account when interpreting the results. The first is that the original HyperKvasir has 23 labeled classes, but 7 classes with less than 100 images were discarded due to their small sample sizes. For this reason, the present system was tested on 16 classes and 10,479 images, and the results of the very rare gastrointestinal findings are not known. Second, the experiments were conducted with one publicly available dataset, so differences in endoscopy equipment, acquisition conditions, institutions and patient populations were not validated by external data. Thirdly, while SADC-Swin was able to outperform the other classifiers in the discrimination of certain hard-to-discriminate classes, there was still some confusion between clinically related classes, specifically between ulcerative-colitis grades and oesophagitis classes. Fourth, the structure-aware auxiliary tasks were specifically tailored to the known relationships in the HyperKvasir label set, which might need to be adapted if used on data sets with different diagnostic taxonomies. Lastly, although the final models were additionally evaluated on an independently reserved official fold, the experiments were conducted on a single public dataset. Further evidence of generalizability would be obtained from full five-fold cross validation and external validation with different institutions, endoscopy systems and patient populations.

7. Conclusion and Future Work

In this study, we proposed a Structure-Aware Dual-View Contrastive Swin Transformer (SADC-Swin) for multi-class gastrointestinal endoscopic image classification. The model is built on top of an ImageNet pre-trained Swin-V2-T backbone

and incorporates dual-view supervised contrastive learning, ordinal supervision for ulcerative-colitis severity, and structure-aware supervision for the normal-Z-line/oesophagitis group. The proposed approach takes advantage of clinically meaningful relations between visually similar classes during training, and uses a standard 16-class classification head during inference, which differs from conventional multiclass learning.

The experimental results showed consistent improvements over the conventional Swin-V2-T baseline. SADC-Swin increased accuracy from 93.06% to 93.87%, Macro-F1 from 87.43% to 89.02%, weighted F1 from 92.75% to 93.76%, balanced accuracy from 87.78% to 89.30%, and MCC from 0.9246 to 0.9334. The proposed structure-aware supervision of difficult and clinically related classes showed the best gains at the class level for ulcerative-colitis grade 1 and oesophagitis-A. Importantly, the performance enhancement was also maintained on the independently reserved fold 0, where the accuracy of SADC-Swin was 93.39%, whereas the baseline was 92.82%, and the Macro-F1 of SADC-Swin was 87.74% while the baseline was 86.11%.

Future work will involve testing SADC-Swin with full five-fold cross validation and external gastrointestinal datasets, generalize the framework to less common HyperKvasir classes, and explore more adaptable methods to learn class relationships instead of manually defining them. Uncertainty estimation and model interpretability could also be added to the further studies to aid the clinical evaluation of unclear endoscopic results.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] H. Ali et al., “Artificial intelligence in gastrointestinal endoscopy: A comprehensive review,” *Annals of Gastroenterology*, vol. 37, no. 2, pp. 133–141, 2024, doi: 10.20524/aog.2024.0861.
- [2] I. S. Seger, “Evaluating and selecting optimal CNN architectures for accurate pneumonia detection in chest X-rays,” *Int. J. Comput. Electron. Aspects Eng.*, vol. 5, no. 4, pp. 183–193, Dec. 2024, doi: 10.26706/ijceae.5.4.20241106. The study compares CNN configurations for medical-image classification.
- [3] H. Borgli et al., “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific Data*, vol. 7, Art. no. 283, 2020, doi: 10.1038/s41597-020-00622-y.
- [4] N. Waisi, “Deep feature fusion method for images classification,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 5, no. 4, pp. 148–153, Dec. 2024, doi: 10.26706/ijceae.5.4.20241103.
- [5] Z. Liu et al., “Swin Transformer V2: Scaling up capacity and resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12009–12019, doi: 10.1109/CVPR52688.2022.01170.
- [6] M. K. Yousif, “Enhanced diabetic retinopathy classification using hybrid EfficientNetB0 with CBAM,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 2, pp. 176–188, May 2026, doi: 10.26706/ijceae.7.2.20260307.
- [7] R. W. Stidham et al., “Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis,” *JAMA Network Open*, vol. 2, no. 5, Art. no. e193963, 2019, doi: 10.1001/jamanetworkopen.2019.3963.
- [8] B. Gutierrez Becker et al., “Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data,” *Therapeutic Advances in Gastrointestinal Endoscopy*, vol. 14, 2021, doi: 10.1177/2631774521990623.
- [9] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18661–18673.
- [10] K. Ramamurthy, T. T. George, Y. Shah, and P. Sasidhar, “A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images,” *Diagnostics*, vol. 12, no. 10, Art. no. 2316, 2022, doi: 10.3390/diagnostics12102316.
- [11] R. Espantaleón-Pérez, I. Jiménez-Velasco, R. Muñoz-Salinas, and M. J. Marín-Jiménez, “Empirical study of attention-based models for automatic classification of gastrointestinal endoscopy images,” in *Computer Analysis of Images and Patterns (CAIP 2023), Lecture Notes in Computer Science*, vol. 14185, 2023, pp. 98–108, doi: 10.1007/978-3-031-44240-7_10.

- [12] W. Wang, X. Yang, and J. Tang, "Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4452–4461, 2023, doi: 10.1109/TCSVT.2023.3277462.
- [13] H. Ge et al., "Development and validation of deep learning models for the multiclassification of reflux esophagitis based on the Los Angeles classification," *Journal of Healthcare Engineering*, vol. 2023, Art. no. 7023731, 2023, doi: 10.1155/2023/7023731.
- [14] H. Guo, S. A. Somayajula, R. Hosseini, and P. Xie, "Improving image classification of gastrointestinal endoscopy using curriculum self-supervised learning," *Scientific Reports*, vol. 14, Art. no. 6100, 2024, doi: 10.1038/s41598-024-53955-8.
- [15] Z. Wang, C. Liu, S. Zhang, and Q. Dou, "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Lecture Notes in Computer Science*. Cham, Switzerland: Springer, 2023, pp. 101–111, doi: 10.1007/978-3-031-43996-4_10.
- [16] J. Lasington and L. S. M. Lasington, "Can deep learning architectures reliably classify the full spectrum of gastrointestinal endoscopic findings? A comparative study of vision transformers and CNNs on 23-class classification," *Gut*, vol. 75, suppl. 2, p. A373, 2026, doi: 10.1136/gutjnl-2026-IDDF.264.
- [17] R. Albasrawi, M. Ilyas, and Q. Althebyan, "Modified YOLOv8x model for coronary stenosis detection and troponin risk stratification," *Discov. Comput.*, vol. 29, Art. no. 34, 2026, doi: 10.1007/s10791-026-09913-1.
- [18] R. Albasrawi, M. Ilyas, and O. Bayat, "A modified YOLO-based deep learning model for accurate ST-elevation myocardial infarction detection in ECG images," *Neural Comput. Appl.*, vol. 38, Art. no. 617, 2026, doi: 10.1007/s00521-026-12347-x.
- [19] A. B. Salih, R. B. Mohammed, A. A. Othman, S. A. Mostafa, and M. Kamalrudin, "YOLOv8-based real-time pedestrian detection model for urban traffic surveillance systems," *Int. J. Inform. Vis.*, vol. 10, no. 2, pp. 556–564, Mar. 2026.
- [20] S. I. Hamad, "Utilizing convolutional neural networks for the identification of lung cancer," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 6, no. 1, pp. 35–41, Mar. 2025, doi: 10.26706/ijceae.6.1.20250206.
- [21] F. Fadhil, M. Abdulghani, A. Salih, and M. Ghazal, "Traffic surveillance: Vehicle detection and pose estimation based on deep learning," *Przegląd Elektrotechniczny*, vol. 99, no. 2, pp. 131–134, 2023, doi: 10.15199/48.2023.02.22.
- [22] N. Waisi, "Comparative evaluation of ResNet-50 and EfficientNet-B1 for pneumonia detection in chest X-ray images using transfer learning," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 6, no. 2, pp. 89–97, Jun. 2025, doi: 10.26706/ijceae.6.2.20250405.
- [23] M. Ghazal, M. Al-Ghadhanfari, and F. F. Fadhil, "Classification of breast cancer using a precise deep learning model architecture," *International Journal of Informatics and Communication Technology*, vol. 14, no. 3, pp. 933–940, 2025, doi: 10.11591/ijict.v14i3.pp933-940.
- [24] M. K. Yousif, "Enhanced diabetic retinopathy classification using hybrid EfficientNetB0 with CBAM," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 2, pp. 176–188, May 2026, doi: 10.26706/ijceae.7.2.20260307