

Article

A Sequential Segmentation and Classification Learning Approach for Skin Lesion Images

Mirco Gallazzi ^{*,†} , Ignazio Gallo ^{*,†}  and Silvia Corchs [†] 

Department of Theoretical and Applied Science, University of Insubria, 21100 Varese, Italy; silvia.corchs@uninsubria.it

* Correspondence: mgallazzi2@uninsubria.it (M.G.); ignazio.gallo@uninsubria.it (I.G.)

† These authors contributed equally to this work.

Abstract

This study investigates how the learning order between segmentation and classification tasks influences performance and generalization in medical image analysis. We propose a Sequential Swin Transformer framework that reuses a shared Transformer backbone with alternating task-specific heads to compare two sequential strategies: (i) segmentation followed by classification and (ii) classification followed by segmentation. Unlike conventional multitask or preprocessing-based pipelines, the proposed framework isolates the impact of task ordering on feature transfer under an identical architecture. Evaluated on the HAM10000 skin lesion dataset, the segmentation-then-classification configuration achieves the highest multiclass accuracy (up to 86.9%) while maintaining strong segmentation performance (Jaccard index $\approx 86\%$). Statistical tests confirm its superiority in accuracy and macro F1 score, whereas Grad-CAM and t-distributed stochastic neighbor embedding (t-SNE) analyses reveal that segmentation-first training yields more lesion-centered attention and a more discriminative latent space. Cross-domain evaluation on gastrointestinal endoscopy images further demonstrates robust segmentation (Jaccard index $\approx 91\%$) and multiclass accuracy ($\approx 94.5\%$), confirming the generalizability of the sequential paradigm. Overall, the proposed method provides a theoretically grounded, clinically interpretable, and reproducible alternative to joint multitask learning approaches, enhancing feature transfer and generalization in medical imaging.



Academic Editor: Vladislav Toronov

Received: 20 October 2025

Revised: 18 November 2025

Accepted: 25 November 2025

Published: 28 November 2025

Citation: Gallazzi, M.; Gallo, I.; Corchs, S. A Sequential Segmentation and Classification Learning Approach for Skin Lesion Images. *Appl. Sci.* **2025**, *15*, 12614. <https://doi.org/10.3390/app152312614>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: deep learning; medical imaging; segmentation and classification; sequential learning; skin lesion analysis; swin transformer; transfer learning; gastrointestinal disease detection

1. Introduction

Skin cancer is highly prevalent worldwide, and earlier diagnosis markedly improves prognosis: five-year survival reaches 95–100% for early-stage melanoma but drops to 25–35% in advanced stages [1–3]. In this context, medical images can be viewed as structured biomedical signals whose spatial patterns encode clinically relevant information.

Deep Learning (DL) has become central to automated skin lesion analysis, providing accurate diagnostic support through data-driven feature learning. Among recent architectures, convolutional neural networks (CNNs) initially dominated the field, achieving strong results in both lesion segmentation and classification tasks. However, their reliance on local receptive fields limits their capacity to capture long-range contextual relationships that are often critical in medical imaging. To overcome these constraints, attention-based

architectures—particularly Transformers—have been introduced. Among them, Swin Transformers (SWIN) have demonstrated excellent performance in capturing both global and fine-grained lesion patterns essential for accurate diagnosis [4].

Despite these advances, reproducibility and generalization remain challenging issues, often restricted by heterogeneous dataset splits and non-standard evaluation protocols. The HAM10000 dataset [5] (HAM) represents a notable exception, offering standardized partitions and an external test set that enable fair and transparent benchmarking in dermatological image analysis.

Segmentation has often been employed as a preprocessing step to assist classification, yet empirical evidence indicates that this approach does not consistently improve diagnostic accuracy [6]. These limitations have motivated the exploration of transfer learning between segmentation and classification tasks [7], seeking to leverage their complementary nature. However, a critical question remains: how does the order in which these tasks are learned affect the transfer of useful representations?

Multitask CNN frameworks, such as the one proposed by He et al. [8], showed that jointly learning segmentation and classification can enhance lesion understanding and diagnostic accuracy. Nevertheless, their convolutional inductive biases restrict the modeling of global dependencies, leaving Transformers better suited for integrating both local and global contextual cues. Yet, it is still unclear whether the sequence in which segmentation and classification are learned influences how Transformers form and transfer such representations. Previous studies that combined these tasks sequentially [9,10] did not explicitly compare training orders, and no systematic analysis has yet been conducted for Transformer-based models.

To address this gap, we propose the Sequential Swin Transformer (SST), a modular framework that features a shared Transformer backbone and task-specific heads, isolating the role of task sequencing. Here, “sequential” refers to a training strategy—not a pipeline where one task’s output feeds the next—where tasks are learned consecutively (starting from segmentation or classification) while reusing the backbone weights. This form of sequential transfer learning, detailed in Section 3, enables us to quantify how ordering affects representation transfer and out-of-distribution robustness—key objectives in biomedical signal interpretation and clinical decision support.

Beyond quantitative metrics, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [11] to qualitatively assess interpretability and evaluate whether different training orders lead the model to focus more precisely on lesion-relevant regions rather than background features.

To evaluate performance and generalizability, we target datasets providing both classification labels and segmentation masks. The HAM dataset [5], with pixel-wise annotations from the ISIC challenge [12], enables consistent within-domain assessment of both tasks. We further probe cross-domain generalization in gastrointestinal imaging using Kvasir [13], which uniquely offers paired labels for classification and segmentation. This evaluation does not aim to compare unrelated clinical contexts but to test the robustness, adaptability, and domain-transfer potential of the proposed sequential learning strategy.

The main contributions of this study are as follows:

- The Sequential Swin Transformer (SST) framework is proposed, supporting both segmentation and classification via task-specific heads and a shared Transformer backbone.
- A systematic comparison is conducted between two sequential configurations: segmentation-then-classification (SST_SC) and classification-then-segmentation (SST_CS), using standardized benchmarks for skin lesion classification and segmentation.

- The framework's generalizability is further assessed through cross-domain evaluation on the gastrointestinal lesion dataset [13], complemented by Grad-CAM analyses to qualitatively examine spatial attention and model interpretability.

The remainder of this paper is organized as follows: Section 2 reviews related work and current limitations. Section 3 describes the proposed framework. Section 4 presents the datasets and experimental results on skin lesion analysis. Section 5 reports the cross-domain generalization results on the gastrointestinal dataset. Finally, Section 6 concludes with a discussion and future research directions.

2. Related Work

Automated skin lesion analysis involves two main tasks: segmentation and classification. The first delineates lesion boundaries, while the second assigns diagnostic categories. Deep learning methods have evolved from convolutional networks (CNNs) to Transformer-based architectures, achieving increasing accuracy and robustness.

Segmentation. Early methods used encoder–decoder CNNs such as DeepLabV3+ [14], which achieved an mIoU of 78.6% on ISIC data [12]. Attention and multi-scale mechanisms were later introduced to handle lesion heterogeneity and low contrast. Masood et al. [15] combined ASSP-based DeepLabV3+ with UNet variants and DenseNet encoders, obtaining 91% Dice and 84% Jaccard across ISIC challenges. CA-Net [16] integrated spatial, channel, and scale attention (92.08% Dice), while FAT-Net [17] embedded Transformer branches into a lightweight CNN for improved contextual reasoning. Transformer-based models such as BAT [18] and SkinSAM [19] further enhanced boundary precision, the latter achieving 88.79% Dice on HAM. Despite these improvements, segmentation remains challenging due to irregular lesion shapes, color variability, and blurred borders—factors limiting cross-dataset generalization.

Classification. Esteva et al. [20] first demonstrated dermatologist-level performance (AUC 0.91) using a CNN on over 130,000 images. Subsequent works integrated attention and feature fusion: Ali et al. [21] reported 91.93% accuracy on HAM, Datta et al. [22] achieved 93.7% with soft attention, and Transformer-based models [23,24] exceeded 90% accuracy on ISIC 2019 [25]. Capsule-based FixCaps [26] reached 96.49% accuracy on HAM using large-kernel convolutions and CBAM attention. Nonetheless, class imbalance, visual similarity, and inconsistent splits still hinder fair comparison [4,27].

Joint or sequential learning. Most methods treat segmentation as a preprocessing step that crops or localizes lesions before classification [28–30], without analyzing architectural integration or training order. He et al. [8] proposed a multitask CNN with a shared encoder and dual decoders, showing gains from joint optimization but leaving the effect of task order unexplored and restricted to CNN-based designs. Coupled learning strategies can be synergistic [31], yet the mechanisms driving such synergy remain unclear.

Recent Transformer-based frameworks have similarly adopted sequential paradigms. Manzoor et al. [32] introduced a two-stage approach in which a VGG16-based U-Net, trained on ISIC2018, segments lesions whose cropped regions are later classified on HAM using EfficientFormerV2 and SwiftFormer. Although this design improves accuracy, the segmentation and classification stages are trained independently, and dataset splits for classification are not explicitly defined, limiting reproducibility and generalization analysis. Himel et al. [33] combined the Segment Anything Model with a ViT classifier, reformulating HAM as a binary benign–malignant task. Despite reporting high Dice and accuracy, the study lacks details on data partitioning and external validation, making it difficult to assess generalization beyond the binary setting. Overall, these Transformer pipelines remain sequential or decoupled, often relying on incomplete or dataset-specific evaluation protocols that obscure the role of task interaction and execution order.

Motivation. Addressing these gaps, our work systematically investigates the order of task execution within a unified Swin Transformer framework, comparing segmentation-first and classification-first configurations to clarify how sequential learning influences feature transfer and downstream performance. Unlike prior dual-stage approaches [32,33], we ensure consistent data splits, shared backbones, and cross-domain validation, providing the first controlled assessment of task ordering. Furthermore, we evaluate cross-domain generalization on the Kvasir dataset [34], which offers paired segmentation and classification labels for gastrointestinal endoscopy [35–37], extending the analysis beyond the dermatological domain.

3. Sequential Learning Framework

This section introduces the Sequential SST framework, designed to evaluate the effect of task ordering in skin lesion analysis. The proposed architecture is modular, enabling the sequential execution of segmentation and classification tasks while maintaining a shared representation space.

3.1. Model Architecture

The SST is designed to isolate the effect of task ordering in medical image analysis. It reuses a single Swin backbone across segmentation and classification, allowing a controlled study of how knowledge flows between tasks when trained sequentially. SST comprises a shared Swin backbone and two lightweight, interchangeable heads for segmentation and classification. Swapping the heads does not alter the backbone; thus, any performance difference stems solely from the training order and representation transfer, rather than from architectural changes.

Formally, given an input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, the backbone (swin_large_patch4_window7_224) [38] divides the image into non-overlapping 4×4 patches, embeds them linearly, and processes them through four hierarchical stages with shifted-window attention. Each stage halves spatial resolution and doubles channel depth, producing a final feature map of size $H' = W' = 7$ with $D = 1536$ channels for 224×224 inputs. Removing the default classification head yields a feature extractor:

$$\mathbf{F} = \mathcal{B}(\mathbf{x}), \quad \mathbf{F} \in \mathbb{R}^{D \times H' \times W'}, \quad (1)$$

where $\mathcal{B}(\cdot)$ includes patch embedding, shifted-window attention, and MLP blocks that produce multi-scale semantic and spatial features shared across tasks.

3.1.1. Segmentation Head

The segmentation head reconstructs spatial detail by upsampling $\mathbf{F}$ back to input resolution through five transposed convolutions that double spatial size and reduce channels as

$$1536 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 1.$$

Each layer is followed by BatchNorm and ReLU, ending with a 1×1 convolution generating a one-channel mask:

$$\mathbf{M} \in \mathbb{R}^{1 \times H \times W}. \quad (2)$$

For each transposed convolution, the output size O is given by

$$O = (I - 1) \cdot S - 2P + K + OP, \quad (3)$$

where $S = 2$ is the stride, P the padding, and K the kernel size. The final mask is resized via bilinear interpolation to match (H, W) .

3.1.2. Classification Head

The classification head processes the same representation. After permuting $[H', W', D] \rightarrow [D, H', W']$ and applying adaptive average pooling and flattening, we obtain the pooled descriptor:

$$\mathbf{f} = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \mathbf{F}_{:,i,j} \in \mathbb{R}^D. \quad (4)$$

A dropout layer precedes two fully connected layers arranged in cascade, corresponding to $\text{Linear}(D \rightarrow C_{\text{multi}})$ and $\text{Linear}(C_{\text{multi}} \rightarrow 2)$.

Multiclass head:

$$\begin{aligned} \mathbf{z}_{\text{multi}} &= \mathbf{W}_{\text{multi}} \mathbf{f} + \mathbf{b}_{\text{multi}}, \quad \mathbf{W}_{\text{multi}} \in \mathbb{R}^{C_{\text{multi}} \times D}, \\ \mathbf{y}_{\text{multi}} &= \text{Softmax}(\mathbf{z}_{\text{multi}}) \in [0, 1]^{C_{\text{multi}}}. \end{aligned} \quad (5)$$

Binary head (cascaded from multiclass logits):

$$\begin{aligned} \mathbf{z}_{\text{bin}} &= \mathbf{W}_{\text{bin}} \mathbf{z}_{\text{multi}} + \mathbf{b}_{\text{bin}}, \quad \mathbf{W}_{\text{bin}} \in \mathbb{R}^{C_{\text{bin}} \times C_{\text{multi}}}, \\ \mathbf{y}_{\text{bin}} &= \text{Softmax}(\mathbf{z}_{\text{bin}}) \in [0, 1]^{C_{\text{bin}}}, \quad C_{\text{bin}} = 2. \end{aligned} \quad (6)$$

Here, C_{multi} is the number of diagnostic categories and $C_{\text{bin}} = 2$ corresponds to benign vs. malignant classification. The cascaded structure mirrors the implemented architecture and supports hierarchical reasoning, where multiclass logits guide binary prediction.

3.2. Task Modularity

SST's modular design enables seamless switching between tasks while preserving backbone weights:

- **SST_SC**: segmentation head $\rightarrow$ classification head;
- **SST_CS**: classification head $\rightarrow$ segmentation head.

This design isolates the effect of training order on feature transfer, as illustrated in Figure 1.

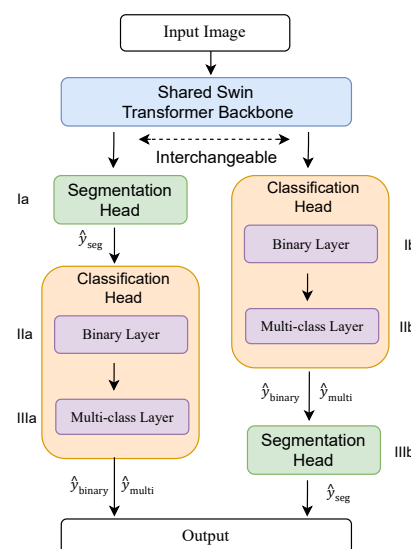


Figure 1. Visual overview of the two sequential training strategies investigated in this study. Left (SST_SC): the model is first trained on segmentation (Ia), then transferred to binary classification (IIa), and finally fine-tuned for multiclass classification (IIIa). Right (SST_CS): training starts with binary classification (Ib), followed by multiclass classification (Iib), and ends with segmentation (IIIb). In both settings, the SWIN backbone is shared across tasks, and task-specific heads are modularly swapped to enable sequential transfer learning.

3.3. Training Pipeline

Two sequential strategies are compared under identical data, backbone, and optimization settings:

- **SST_SC**: segmentation → binary classification → multiclass fine-tuning;
- **SST_CS**: binary classification → multiclass fine-tuning → segmentation.

At each stage, the best validation checkpoint (IoU or accuracy) initializes the next.

SST_SC. The model first learns lesion boundaries using Binary Cross-Entropy (BCE) for masks with Adam optimizer. The checkpoint with the highest validation IoU is retained, then used to initialize the binary classification stage, optimized with two-class Cross-Entropy (consistent with the 2-way Softmax). A final fine-tuning phase performs multiclass classification with Cross-Entropy loss, updating all weights. This sequence assesses whether spatial localization enhances later discriminative learning.

SST_CS. Training starts with binary classification (two-class Cross-Entropy), proceeds to multiclass fine-tuning (Cross-Entropy), and concludes with segmentation (BCE). Each stage reuses the best-performing checkpoint from the previous one, ensuring consistent knowledge transfer.

Joint loss during classification. When both outputs are active (e.g., during multiclass training), the total loss is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{multi}} \mathcal{L}_{\text{multi}} + \lambda_{\text{bin}} \mathcal{L}_{\text{bin}}, \quad (7)$$

where $\mathcal{L}_{\text{multi}}$ and $\mathcal{L}_{\text{bin}}$ are Cross-Entropy losses for multiclass and binary predictions, respectively, and $\lambda_{\text{multi}}, \lambda_{\text{bin}}$ are fixed balancing coefficients. At stages where only one head is optimized, the inactive output is forward-propagated but excluded from gradient computation.

Model selection and metrics. Model selection is metric-driven: the checkpoint with the highest validation IoU is retained for segmentation, and the highest validation accuracy for classification. The adopted metrics are

$$\text{Accuracy}_{\text{task}} = \frac{\text{Correct Predictions}}{\text{Total Samples}} \quad (8)$$

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}} \quad (9)$$

$$\text{Dice} = \frac{2 \cdot \text{Intersection}}{\text{Prediction Size} + \text{Ground Truth Size}} \quad (10)$$

Unlike traditional multitask learning—which optimizes tasks jointly and often suffers from gradient interference—SST optimizes them sequentially, allowing each task to shape the shared backbone independently. This design isolates the causal role of task order in feature transfer and reflects realistic clinical workflows, where localization and diagnosis occur in sequence.

3.4. Implementation Details

Experiments use PyTorch 2.1.0 (CUDA 11.8). Inputs are resized to 224×224 ; images are normalized with ImageNet mean/std. **Data partitioning** was performed randomly with subject-level independence, ensuring that all images from the same patient (including multi-view acquisitions of the same lesion) were assigned to a single subset. Segmentation masks were consistently paired with their corresponding image IDs across all splits. Patient-level metadata from HAM were used to enforce this grouping. No image-level deduplication or near-duplicate filtering was performed, as multi-view samples represent natural clinical variability and were intentionally preserved to enrich the diversity of the training set. Since the external HAMt test set is fully independent, duplicates do not affect external evaluation; potential optimism could arise only in internal splits, but subject-level grouping

prevents this scenario. Metadata regarding acquisition site or device were not available and could not be used for stratification. **Training augmentation** was applied stochastically to each image at every epoch, ensuring that the model never sees the exact same sample twice. Transformations included random resize (shortest side in $[224, 280]$), random crop 224×224 , horizontal flip (prob. 0.5), affine translation up to 10% (prob. 0.5), and rotation in $[-180^\circ, 180^\circ]$ (prob. 0.99). This dynamic augmentation strategy substantially increases data diversity, mitigates class imbalance, and improves robustness to variations in scale, orientation, and illumination, thereby enhancing adaptability to real-world clinical imaging conditions. **Validation** uses deterministic resize to 224×280 and center crop 224×224 . **Test** resizes to 224×224 without cropping. Unless stated otherwise, we use Adam with learning rate $1 \cdot 10^{-4}$ and a LambdaLR scheduler with exponential decay 0.95^e at epoch e . The best checkpoints (as above) initialize subsequent stages. Additional hyperparameters (batch size, epochs) are in Section 4.

3.5. Grad-CAM Protocol

To qualitatively assess the interpretability of the proposed framework, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [11] to visualize the image regions that influence the model's predictions the most. This method provides post hoc spatial attention maps, highlighting whether the model focuses on lesion-relevant areas rather than background regions.

In this study, Grad-CAM is applied to the SST framework to analyze how task ordering affects the learned spatial representations in the two sequential configurations previously described (SST_SC and SST_CS).

Since the main goal is to explore class-discriminative behavior, Grad-CAM is applied exclusively to the multiclass classification head. For each input image, the feature maps $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ are extracted from the Swin backbone, and gradients are computed with respect to the target class logit y_t . The channel-wise importance weights are then calculated as

$$\alpha_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \frac{\partial y_t}{\partial \mathbf{F}_{c,i,j}}, \quad (11)$$

and the Grad-CAM heatmap is obtained as

$$\mathbf{H}_t = \text{ReLU} \left(\sum_{c=1}^C \alpha_c \mathbf{F}_c \right), \quad (12)$$

where $\mathbf{F}_c$ denotes the c -th feature channel of the backbone output, α_c is the corresponding channel weight obtained from Equation (11), and $\text{ReLU}(\cdot)$ ensures that only positive class contributions are retained. The resulting activation map $\mathbf{H}_t$ is normalized to the $[0, 1]$ range, bilinearly upsampled to match the input resolution, and superimposed on the original image to visualize attention intensity.

To ensure consistent interpretation, Grad-CAM is computed using a memory-efficient implementation adapted for Transformer backbones, with multi-crop smoothing (ten random crops per image) to reduce variability.

Qualitative analyses are performed on the HAM dataset. These visualizations complement the quantitative evaluation in Section 4, providing insight into how task order influences the model's spatial attention.

3.6. Centroid-Based Feature Analysis

For the multiclass head, we extract penultimate-layer features on the validation set, compute class-wise centroids, and measure (i) the mean inter-class centroid distance and

(ii) the minimum pairwise centroid distance. We report values at representative checkpoints to quantify the temporal evolution of the feature space.

4. Experiments and Results

This section describes the datasets, experimental setups, and evaluation metrics used to assess how task ordering in the proposed SST framework affects segmentation and classification performance under the STL paradigm.

4.1. Datasets

Experiments are conducted on the ISIC 2018 Challenge dataset [39], which is primarily based on the HAM collection [5]. It includes 10,015 dermoscopic images categorized into seven diagnostic classes: MEL, NV, BCC, AKIEC, BKL, DF, and VASC (Figure 2 illustrates a representative lesion example for each class). The official training, validation, and test splits provided by the ISIC 2018 organizers are adopted for both segmentation and classification tasks. Class imbalance remains substantial, with NV accounting for more than 60% of the samples—reflecting clinical prevalence but posing challenges for multiclass training.

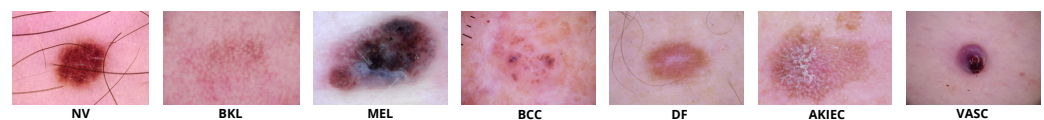


Figure 2. Representative skin lesion images from the HAM dataset, annotated with their respective diagnostic classes. In order, from left to right, we have Nevus (NV), Benign Keratosis-like Lesions (BKL), Melanoma (MEL), Basal Cell Carcinoma (BCC), Dermatofibroma (DF), Actinic Keratosis/Intraepithelial Carcinoma (AKIEC), and Vascular Lesions (VASC).

For segmentation, we rely on the subset of 2594 images with pixel-wise lesion annotations from ISIC 2018 Task 1, reporting results on the official 1000-image test set. For classification generalization, the HAMt dataset [40], released in 2023 and containing 1512 previously unseen dermoscopic images, is employed exclusively as an external test set for models trained on ISIC 2018. Together, ISIC 2018 (for training and internal evaluation) and HAMt (for external testing) provide standardized and complementary benchmarks for assessing the generalizability of the proposed sequential learning framework. All datasets used in this study are publicly available and were collected and anonymized by their original authors, following the appropriate ethical approval. The ISIC and HAM collections are released for research purposes under open data licenses. However, it is important to acknowledge that these datasets are not demographically balanced, with lighter skin tones being predominant, which may limit the generalizability of trained models across diverse populations. Skin-tone or demographic labels are not provided in these datasets, and performance could therefore not be stratified by subgroup. As a result, potential disparities across skin tones may exist but cannot be quantified with the available data.

4.2. Experimental Settings

Two sequential configurations are evaluated: (i) segmentation followed by classification (binary and then multiclass), and (ii) classification followed by segmentation. Each is tested under two conditions.

In the first, models are trained, validated, and tested on HAM (75%/15%/10%) using stratified random sampling to preserve class balance. In the second, models are trained and validated on the official HAM division (80%/20%) and evaluated on the external HAMt test set. Segmentation performance is consistently computed on the 1000 fixed ISIC 2018 test images to ensure reproducibility.

To simplify notation, we define the following configurations:

- **Our_A**: Segmentation → Classification (HAM internal split);
- **Our_B**: Classification → Segmentation (HAM internal split);
- **Our_C**: Classification → Segmentation (trained/validated on HAM, tested on HAMt);
- **Our_D**: Segmentation → Classification (trained/validated on HAM, tested on HAMt).

All models share the same architecture (Swin backbone with task-specific heads) and hyperparameters: input resolution 224×224 , learning rate 1×10^{-4} , and 100 training epochs. The batch size was 144 for the segmentation-first and 128 for the classification-first setups, chosen empirically to optimize GPU utilization. Each experiment was repeated five times, and we report the mean and standard deviation of all metrics to ensure statistical reliability.

4.3. Evaluation Metrics

Segmentation performance is assessed using the IoU (9) and Dice coefficients (10). For classification, both binary and multiclass tasks are evaluated using standard metrics:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{\text{Total Samples}}, & \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}, & \text{F1} &= 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \quad (13)$$

We further report the Area Under the ROC Curve (AUC) using a one-vs.-all approach. Both macro- and weighted-averaged AUC values are computed to account for class imbalance and to evaluate per-class discrimination.

In the following subsections, we present and analyze the results for all configurations (Our_A–Our_D), discussing how task order, dataset partitioning, and inter-task transfer affect segmentation and classification performance.

4.4. Quantitative Evaluation and Analysis

Table 1 summarizes segmentation and classification results across the four experimental setups. We analyze how task order and dataset composition affect performance. In addition, ROC curves and confusion matrices provide class-wise insights.

Table 1. Benchmark results of the SST model on the HAM dataset. TA denotes test accuracy, with the Dataset column distinguishing between HAM (full dataset) and HAMt (external test set). Jaccard and Dice scores evaluate segmentation performance, while TAb and TAm report binary and multiclass classification accuracy, respectively. TPm, TRm, and TF1m correspond to multiclass precision, recall, and F1 score. All values are averaged over five runs.

Author	Dataset	Jaccard (%)	Dice (%)	TAb (%)	TAm (%)	TPm (%)	TRm (%)	TF1m (%)
Our_A	HAM	86.13 ± 0.27	89.61 ± 0.15	93.23 ± 0.59	92.16 ± 0.69	89.75 ± 0.98	84.90 ± 0.89	86.44 ± 1.13
Our_B	HAM	86.03 ± 0.18	89.48 ± 0.22	93.76 ± 0.29	91.94 ± 0.49	89.54 ± 1.40	86.45 ± 1.14	87.68 ± 1.24
Our_C	HAM+Ht	86.58 ± 0.19	89.50 ± 0.25	90.80 ± 0.55	86.63 ± 0.39	82.37 ± 0.51	79.42 ± 0.30	79.74 ± 1.80
Our_D	HAM+Ht	86.26 ± 0.16	89.48 ± 0.23	91.07 ± 0.52	86.87 ± 0.45	84.94 ± 0.69	78.47 ± 1.48	80.57 ± 0.10

Segmentation. Jaccard and Dice are stable across orders. On HAM, Our_A (SST_SC) and Our_B (SST_CS) are nearly identical (Jaccard $\approx 86.1\%$ vs. 86.0%). With the external setting, Our_C (SST_CS) is marginally higher than Our_D (SST_SC) (Jaccard 86.58% vs. 86.26%), indicating that classification may provide features that slightly aid subsequent segmentation. Qualitative examples in Figure 3 confirm sharp and well-localized masks in both orders, corroborating the quantitative stability.

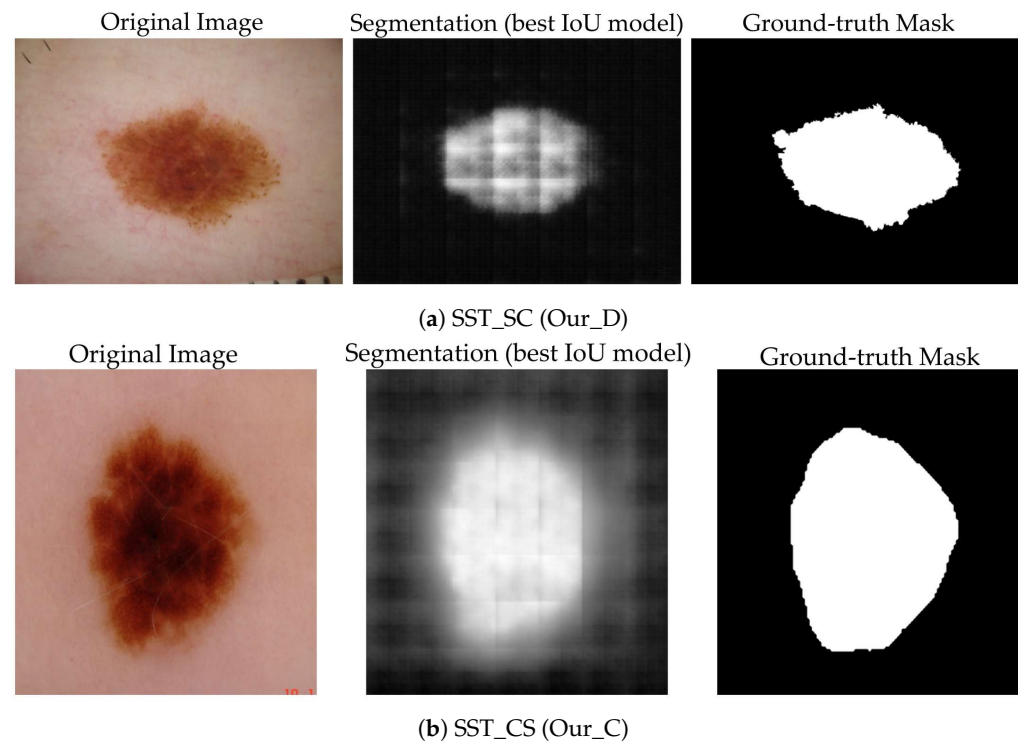


Figure 3. Visual comparison of segmentation results for a representative lesion from the test set, evaluated under two sequential learning configurations. Each row presents, from left to right, the original input image, the predicted segmentation map from the best-performing model, and the ground-truth mask. (a) SST_SC (Our_D): segmentation followed by classification. (b) SST_CS (Our_C): classification followed by segmentation.

Classification. Task order has a clearer impact on multiclass accuracy, especially under domain shift. On HAMt, Our_D (SST_SC) achieves the best TAm ($86.87\% \pm 0.39$) compared to Our_C (SST_CS, $86.63\% \pm 0.36$), with corresponding macro F1 of $80.57\% \pm 0.10$ vs. $79.74\% \pm 1.80$. A one-sided Wilcoxon signed-rank test over five runs supports the advantage of SST_SC: for accuracy, $W = 15.0$, $p = 0.03125$; for macro F1, $W = 15.0$, $p = 0.03125$. Importantly, these tests were computed across independent runs (i.e., across random seeds) rather than across individual predictions, so statistical significance reflects experiment-level variability. Figures 4 and 5 show stronger class-wise ROC behavior for SST_SC, notably on challenging classes (e.g., MEL, AKIEC), with fewer malignant false positives in external evaluation.

Internal vs. external evaluation. Performance drops from HAM to HAMt highlight optimistic bias in internal-only tests. Focusing statistics on Our_C and Our_D (HAMt) provides a more reliable estimate of generalization and isolates the effect of ordering under realistic shift.

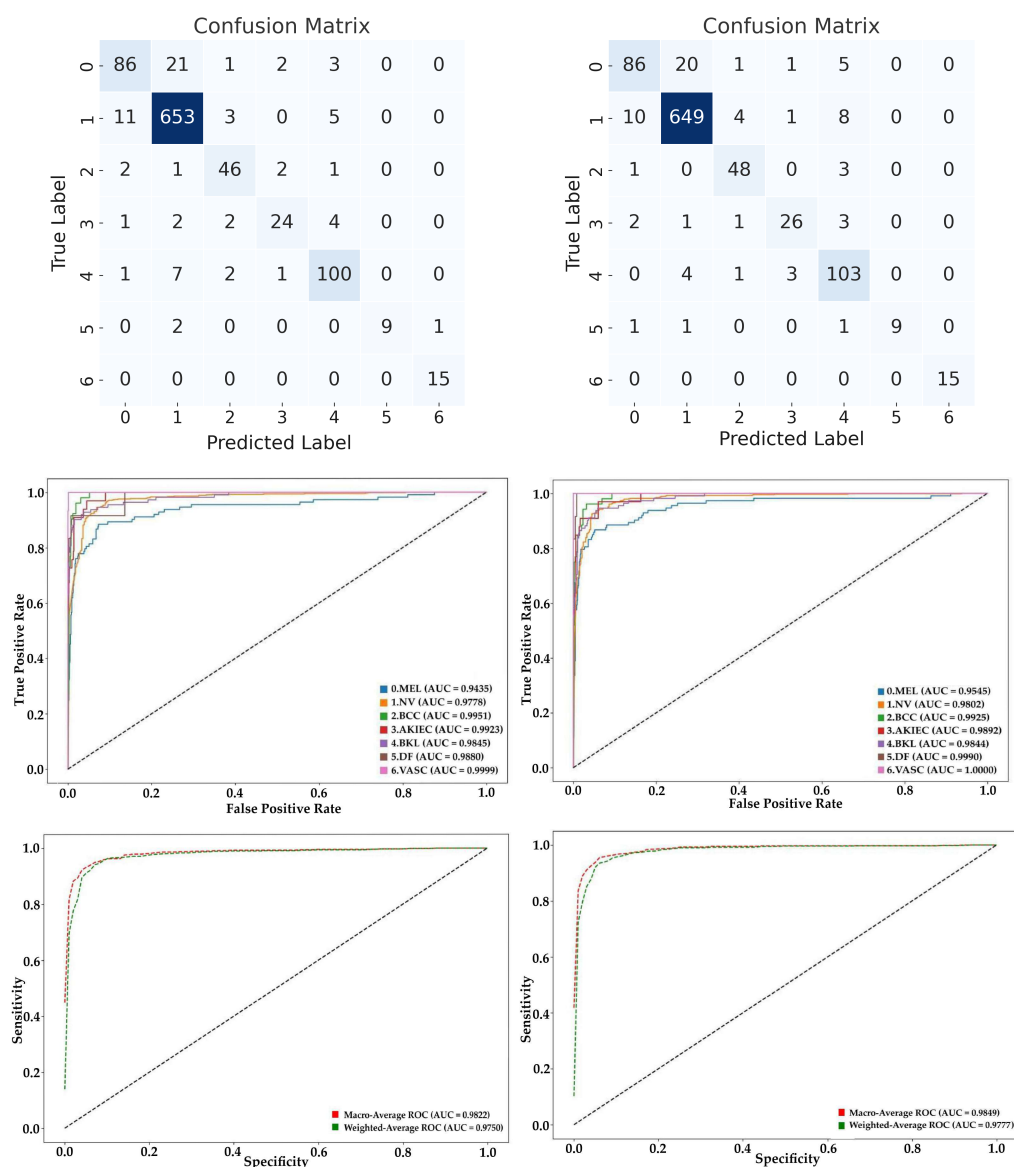


Figure 4. Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (top to bottom) for the models Our_A (left) and Our_B (right), as reported in Table 1. Each row corresponds to an experimental setup using the best validation checkpoint. In the confusion matrices (top), the true labels are on the vertical axis and the predicted labels on the horizontal axis; class indices 0–6 correspond to MEL, NV, BCC, AKIEC, BKL, DF, and VASC. The middle row displays per-class ROC curves, showing True Positive Rate (TPR, sensitivity) versus False Positive Rate (FPR), while the bottom row reports macro and weighted average ROC curves, plotting sensitivity against specificity.

4.5. Qualitative Analysis via Grad-CAM

Grad-CAM maps (Figure 6) for the multiclass head provide a qualitative view of how task ordering shapes spatial attention. In the SST_SC configuration, activations are consistently centered on the lesion core and conform to its morphological boundaries and texture, while SST_CS shows broader and less discriminative responses extending into peri-lesional skin. These observations highlight that learning segmentation first promotes spatially constrained and pathology-relevant activations.

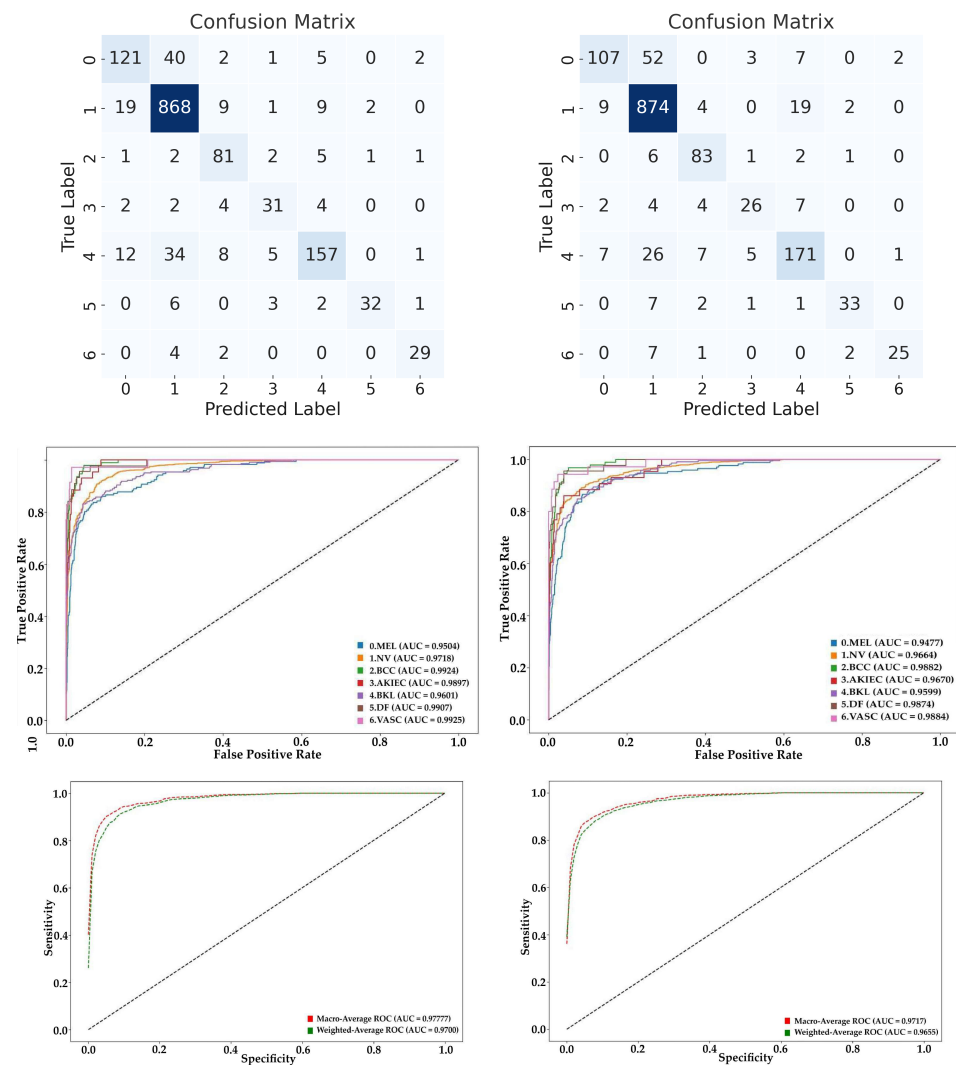


Figure 5. Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (top to bottom) for models Our_C (left) and Our_D (right), as reported in Table 1. Each row represents an experimental setup using the best validation checkpoint. In the confusion matrices (top), the true labels are on the vertical axis and the predicted labels on the horizontal axis; classes 0–6 correspond to MEL, NV, BCC, AKIEC, BKL, DF, and VASC. The middle row displays per-class ROC curves, showing True Positive Rate (TPR, sensitivity) versus False Positive Rate (FPR), while the bottom row reports macro and weighted average ROC curves, plotting sensitivity against specificity.

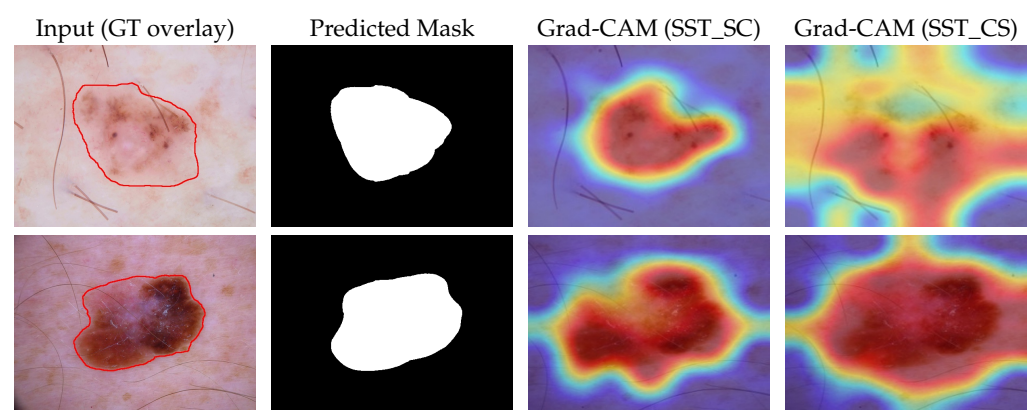


Figure 6. Grad-CAM visualizations on two representative HAM cases comparing SST_SC and SST_CS. From left to right: input image with the ground-truth mask overlaid in red for visualization purposes, predicted binary mask, and Grad-CAM maps for the two sequential configurations.

Two effects are evident: (i) focus tightening, where SST_SC concentrates saliency within lesion areas; (ii) background spillover reduction, with fewer activations outside the lesion contour compared to SST_CS. These qualitative patterns align with the external performance gains in Table 1 and the improved class-wise ROC behavior in Figures 4 and 5.

Figure 7 illustrates the temporal evolution of Grad-CAM activations across training epochs for two representative lesions. The maps show how saliency progressively contracts toward the diagnostically relevant core, confirming that spatial priors learned through segmentation guide and stabilize feature learning during classification. The final epoch reveals compact, lesion-centered activations, providing visual evidence of effective task-to-task transfer.

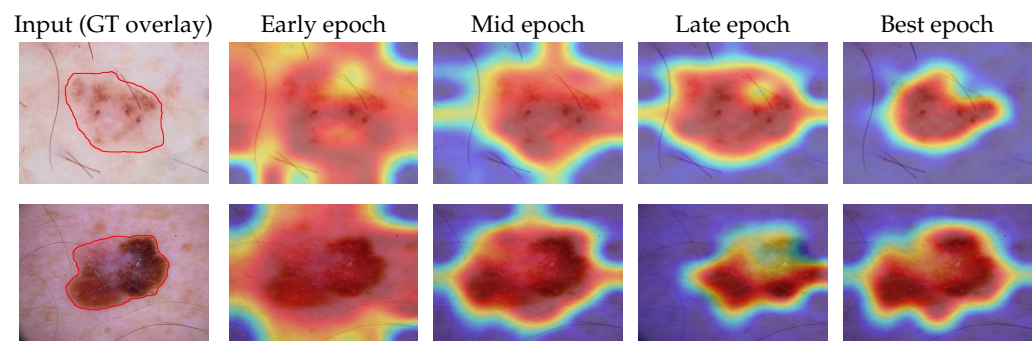


Figure 7. Temporal evolution of Grad-CAM activations for two representative HAM cases shown in Figure 6. Each row corresponds to one lesion. From left to right: input with ground-truth mask overlay, and Grad-CAM maps extracted at early, mid, late, and best training epochs for SST_SC configuration.

Residual activations occasionally persist around illumination artifacts or texture-rich regions, suggesting that part of the network still captures superficial cues. Incorporating stronger spatial constraints—e.g., boundary-aware decoders or adaptive gating—could further mitigate this effect. Overall, these results confirm that segmentation-first training enhances interpretability and offers a clear mechanistic explanation for the superiority of SST_SC on HAMt.

These attention dynamics are consistent with the latent-space analysis presented next, where t-SNE projections show how refined, lesion-focused activations evolve into more compact and discriminative class clusters.

4.6. Feature Space Visualization

We analyze representation geometry via t-SNE projections of the penultimate multiclass layer on HAMt (Figure 8). Compared with a Swin baseline trained only for classification, the sequential model (SST_CS, Our_C in the Figure 8) exhibits tighter clusters and wider inter-class margins, with reduced overlap among visually similar classes (NV, MEL, BKL). Although t-SNE is non-metric and depends on perplexity and initialization, the qualitative separation is consistent across multiple runs and mirrors the external accuracy/macro-F1 trends in Table 1.

We further inspect the temporal evolution of embeddings at checkpoints with improving validation accuracy (Figure 9). Clusters progressively exhibit reduced intra-class distance and increased inter-class separation, indicating that sequential transfer promotes a more structured and discriminative latent space as training progresses. This effect is most visible for classes with initially entangled boundaries (e.g., MEL vs. BKL), suggesting that spatial priors learned earlier help disentangle borderline phenotypes.

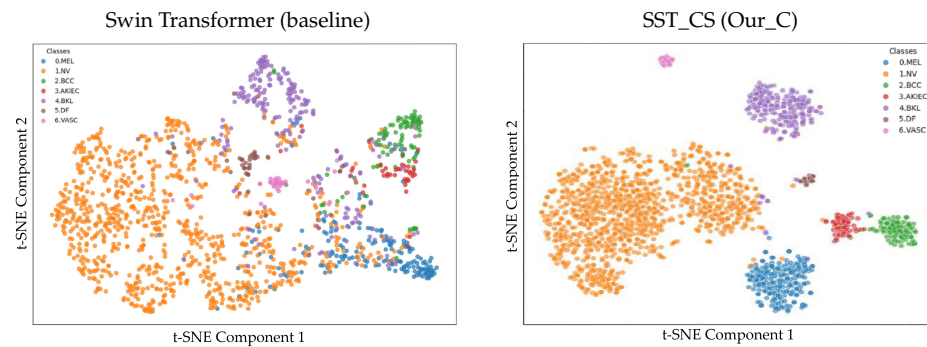


Figure 8. t-SNE projections of the latent features on the external HAMt test set. **(Left)** Swin Transformer trained for classification only. **(Right)** Sequential Swin Transformer (SST) trained with the classification–segmentation order. Each color corresponds to one of the seven diagnostic classes.

Two caveats apply. First, class imbalance can still induce partial overlap, even under sequential transfer, as reflected by the remaining confusion in ROC and F1 for difficult classes. Second, t-SNE provides a qualitative view; future work could complement it with quantitative alignment metrics (e.g., class-wise centroid distances, between- and within-scatter) or representation similarity analysis to more rigorously assess how ordering alters feature geometry. Despite these caveats, the observed trajectories align with the superiority of segmentation-first under domain shift.

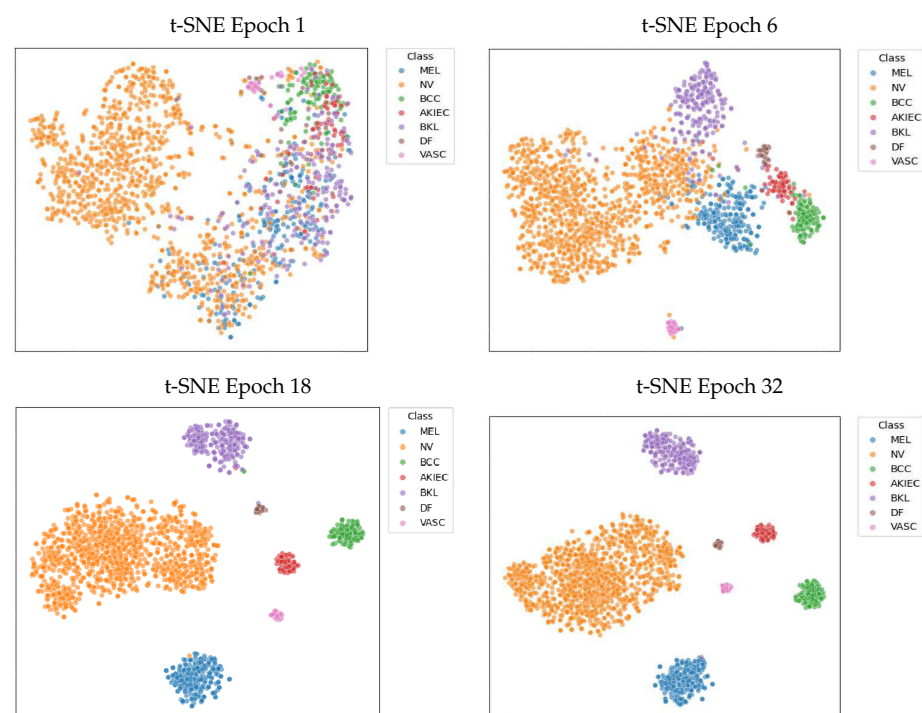


Figure 9. t-SNE projections of multiclass feature embeddings on the HAMt validation set at different training stages of the SST model. As training progresses, clusters become more compact and inter-class separation increases, confirming the effectiveness of sequential transfer learning in shaping the latent space.

4.7. Centroid-Based Quantification of Feature Space

To complement the qualitative analysis based on t-SNE, a quantitative evaluation of the latent feature geometry through class-wise centroid distances has been introduced. For each validation checkpoint, the feature vectors from the penultimate layer of the multiclass head are collected and grouped by class. The centroid of each class is then computed as

the mean of its feature vectors, and pairwise Euclidean distances between centroids are used to measure both (i) the mean inter-class separation and (ii) the minimum pairwise distance across all class pairs. These metrics quantify how the shared representation evolves as training progresses and reveal how the segmentation-first sequence shapes a more discriminative embedding space.

As reported in Table 2, the mean inter-class centroid distance increases from 28.64 (epoch 1) to 53.42 (epoch 32), indicating progressively wider margins between lesion categories. Similarly, the minimum distance between the closest class centroids rises from 8.05 to 22.57, suggesting that classes with overlapping characteristics become more separable over time. These results quantitatively substantiate the qualitative observations obtained through t-SNE and Grad-CAM analyses, showing that early segmentation training drives the model toward a more structured and discriminative feature space.

Table 2. Centroid-based quantification for the segmentation-followed-by-classification configuration. The reported distances are computed on the validation set at representative training epochs.

Epoch	Mean Inter-Class Distance	Closest Class Pair	Minimum Distance
1	28.64	AKIEC–DF	8.05
6	44.98	AKIEC–BCC	13.97
18	47.01	AKIEC–BCC	18.50
32	53.42	BKL–DF	22.57

4.8. Comparison with Literature

To evaluate the effectiveness of our SST framework, we benchmark it against state-of-the-art (SOTA) methods in both segmentation and classification on the HAM dataset.

Segmentation. Table 3 shows that the proposed SST achieves highly competitive segmentation performance on HAM. Both Our_A and Our_B configurations reach a Jaccard index around 86%, surpassing traditional convolutional architectures such as GAN-UNET (77.0%) and DeepLabV3+ (82.8%). Although Transformer-based or hybrid methods like BAT [18], CA-Net [16], and the Polar transformation approach [41] achieve slightly higher Dice scores (up to 92.5%), SST reduces the gap to within 1–2 percentage points despite not being specifically optimized for segmentation. This demonstrates that the unified Swin backbone can provide robust localization capabilities while maintaining architectural simplicity. When trained on the combined HAM+HAMt dataset (Our_C/Our_D), the results show a slight gain in both Jaccard and Dice scores, suggesting that additional data diversity enhances generalization stability. Overall, the SST framework achieves near SOTA segmentation performance without requiring task-specific tuning.

Table 3. Segmentation benchmark on the HAM dataset. All values are percentages. Asterisks (*) denote values as originally reported. The Jaccard index refers to the Intersection over Union (IoU) metric.

Method	Jaccard (%)	Dice (%)
GAN-UNET [15]	77.00	85.20
SkinSAM [19]	78.43 *	88.79 *
DenseNet121-UNET [15]	83.50	89.70
DeepLabV3+ [15]	82.80	89.80
CA-Net [16]	–	92.08
BAT [18]	84.30	92.10
Polar Image Transformation [41]	87.43	92.53
SST_SC (Our_A)	86.13 ± 0.27	89.61 ± 0.15
SST_CS (Our_B)	86.03 ± 0.18	89.48 ± 0.22
SST_CS (HAM+Ht, Our_C)	86.58 ± 0.19	89.50 ± 0.25
SST_SC (HAM+Ht, Our_D)	86.26 ± 0.16	89.48 ± 0.23

Classification. From Table 4, SST_SC achieves 92.16% accuracy on the internal HAM split and 86.87% on the external HAMt test set. At first glance, these figures may appear modest compared with the $> 96\%$ accuracy reported by methods such as FixCaps [26] or Medical Vision Transformers [42]. A broader comparison, including recent Transformer-based and multitask architectures such as the SAM-ViT framework [33] and the dual-stage VGG16-UNet pipeline [32], further confirms the competitiveness of the proposed SST. While these approaches report comparable internal accuracies ($\approx 91\text{--}93\%$), they often rely on domain-specific or non-standard evaluation settings, making direct cross-study comparison challenging. However, direct comparison is not straightforward: most studies rely on non-standard or undocumented splits, often evaluating on data highly similar to the training set. When FixCaps is re-evaluated under our same external protocol (Exp1–3), performance drops to 75–76%, a decline of roughly 20 percentage points. This highlights the difficulty of reproducing such high scores and underscores the importance of fair, external testing for assessing true generalization. Under these standardized conditions, the proposed SST ranks among the strongest and most reproducible models, demonstrating stable accuracy and F1 scores across domains while maintaining a transparent and unified evaluation framework.

Table 4. Multiclass classification benchmark results on the HAM dataset (HAM_b stay for the use of HAM as a binary dataset). The “Split” column indicates the dataset partitioning into training, validation, and test sets. For entries with parentheses (e.g., “80 (90/10) – 20”), the portion inside represents the training/validation split, while the value outside refers to the test set. An asterisk (*) indicates that no test set was used or reported. TA_multiclass (%) reports the accuracy for multiclass classification.

Author	Method	Dataset	Split	TA_Multiclass (%)
Mushtaq et al. [43]	Ensemble VGG16	HAM	80–20 + 15% not duplicate in Test	89
Jain et al. [44]	TL on Xception	HAM	80 (90–10) – 20	89.66
Himel et al. [33]	SAM + ViT (binary–multiclass transfer)	HAM_bin	80–20 *	91.80
Chaturvedi et al. [45]	InceptionV3	HAM	88–12 *	91.56
Manzoor et al. [32]	VGG16-UNet + EfficientFormerV2/ SwiftFormer	HAM	80–20 *	92.50
Chaturvedi et al. [45]	InceptionResNetV2	HAM	88–12 *	93.20
Shetty et al. [46]	CNN models	HAM	80–20	95.18
Aladhadh et al. [42]	Medical Vision Transformer	HAM	70–20–10	96.14
Lan et al. [26]	FixCaps	HAM	85–15	96.49
Exp1	FixCaps	HAM+HAMt	80 + HAMt	76.19
Exp2	FixCaps	HAM+HAMt	80 + HAMt	75.28
Exp3	FixCaps	HAM+HAMt	80 + HAMt	75.60
Our_A	SST_SC	HAM	75–15–10	92.16 $\pm$ 0.69
Our_B	SST_CS	HAM	75–15–10	91.94 $\pm$ 0.49
Gallazzi et al. [4]	Swin Transformer	Large Dataset	80–20 + HAMt	86.37
Our_C	SST_CS	HAM+HAMt	80–20 + HAMt	86.63 $\pm$ 0.39
Our_D	SST_SC	HAM+HAMt	80–20 + HAMt	86.87 $\pm$ 0.45

Against our prior preprocessing pipeline. Compared with the preprocessing-based approach in [6], the sequential SST consistently improves both segmentation and multiclass classification (Table 5). Notably, Our_D reaches 86.87% on HAMt vs. 84.13% in the previous best pipeline, confirming the benefit of learning shared features and transferring them sequentially.

Table 5. Comparative results between our previous preprocessing-based pipeline [6]—which used segmentation outputs as inputs to classification—and the proposed sequential SST models trained on HAM and tested on HAMt. The first three rows report the baseline strategy, while SST_CS and SST_SC represent the joint sequential learning configurations from Table 1. Metrics include segmentation performance (Jaccard %) and multiclass classification results: accuracy (TAm %), precision (TPm), recall (TRm), and F1 score (TF1m).

Author	Method	Jaccard (%)	TAm (%)	TPm (%)	TRm (%)	TF1m (%)
Gallazzi et al. [6]	HAM+YOLO	77.00	84.01	84.09	84.12	83.53
Gallazzi et al. [6]	HAM+DeepLabV3	81.66	84.12	84.38	84.12	83.75
Gallazzi et al. [6]	HAM+ST	82.75	84.13	84.41	84.12	83.58
Our_C	SST_CS	86.58 ± 0.19	86.63 ± 0.39	82.37 ± 0.51	79.42 ± 0.30	79.74 ± 1.80
Our_D	SST_SC	86.26 ± 0.16	86.87 ± 0.45	84.94 ± 0.69	78.47 ± 1.48	80.57 ± 0.10

4.9. Ablation and Interpretative Analysis

Unlike conventional ablations that modify hyperparameters or the depth of the architecture, our analysis focuses on understanding the interplay between data availability, experimental reproducibility, and the qualitative and quantitative evidence gathered through statistical and feature space analyses.

Dataset constraints and reproducibility. Ensuring fair comparisons in medical image analysis remains a major challenge. Public datasets often vary in structure, labeling accuracy, and predefined splits, making reproducibility across studies challenging and potentially leading to biased results. Smaller datasets, such as ISIC-2017 or ISIC-2016 [47], include both segmentation masks and classification labels but are too limited in size for reliable multiclass evaluation. These discrepancies hinder direct cross-study comparison and emphasize the need for standardized, transparent, and reproducible evaluation protocols. To address this issue, our work adopts fixed and publicly available test sets for both segmentation and classification, ensuring consistent benchmarking and enabling fair performance assessment across models. Furthermore, the complete training configuration, including epochs, learning rate, and augmentation policies, is standardized and reported earlier to facilitate reproducibility and support deployment-oriented Sections 3.3 and 4.

Sequential order and statistical validation. A joint multitask baseline with a shared backbone and simultaneous optimization of segmentation and classification was not included in this study. Our focus was to isolate the causal effect of task ordering under strictly identical architectural and optimization conditions; nonetheless, we acknowledge that a well-tuned multitask baseline would constitute a valuable point of comparison and represents a limitation of the current work.

As shown in Section 4.4, SST_SC significantly outperforms SST_CS on HAMt multiclass accuracy and macro F1, with the Wilcoxon signed-rank test ($W = 15.0$, $p = 0.03125$) computed across independent runs (i.e., across random seeds) rather than across individual predictions, ensuring that statistical significance reflects experiment-level variability, and supporting the hypothesis that spatial priors learned first enhance downstream discrimination.

Interpreting Grad-CAM and t-SNE results. The Grad-CAM analysis (Section 4.5) provides clear qualitative evidence of the benefits of learning segmentation before classification. In the SST_SC configuration, attention maps consistently concentrate within the lesion core, accurately aligning with diagnostically meaningful regions and capturing the morphological structure of the lesion. In contrast, SST_CS exhibits broader and less discriminative activation patterns that often extend into surrounding healthy skin, indicating weaker localization of disease-specific features. These findings confirm that segmentation-first training explicitly guides the model toward pathology-relevant cues, enhancing interpretability and offering a transparent understanding of what the network attends to during

decision-making. Nonetheless, residual activation zones unrelated to pathological areas can still be observed. These secondary responses suggest that a portion of the learned attention continues to capture superficial texture variations rather than purely diagnostic features. This behavior corresponds to the partial class overlaps identified in the t-SNE projections (Figure 8), where embeddings of visually similar categories—such as MEL and BKL—remain partially entangled. Together, these qualitative and latent-space observations demonstrate that while segmentation-first learning significantly improves spatial focus and interpretability, further refinement is needed to suppress residual background activations and fully disentangle feature representations across closely related classes.

Implications for segmentation-head design. Since the segmentation decoder shapes the spatial priors transferred to classification, refining its design (e.g., richer decoders, boundary-aware modules, adaptive gating) is a promising direction for further stabilizing attention and filtering irrelevant cues before they propagate to the classifier. Building on these analyses, the following discussion extends the comparison to recent Transformer-based approaches and outlines the broader interpretative and clinical implications of sequential learning.

Comparison with prior Transformer-based frameworks. In contrast to recent Transformer-based dual-stage pipelines [32,33], which sequentially applied segmentation and classification without controlling for dataset overlap or optimization consistency, the SST framework isolates task ordering as the only varying factor within a unified backbone. This controlled formulation eliminates cross-dataset bias and training dependency, enabling a direct and reproducible assessment of how segmentation-first learning reshapes the downstream feature space and classification behavior.

Ethical and clinical interpretability considerations. Beyond quantitative improvements, interpretability plays a central role in establishing model reliability and clinical usability. The transparent behavior observed through Grad-CAM facilitates expert verification, ensuring that the network’s focus corresponds to diagnostically meaningful regions and enabling the identification of potential biases or spurious correlations. Building upon this interpretability layer, sequential frameworks such as SST could be further coupled with multimodal [48] or large language models to combine visual reasoning with contextual clinical knowledge [49,50], ultimately improving diagnostic accuracy, traceability, and decision support in real-world dermatology workflows. Recent evaluations have also shown that general-purpose, non-fine-tuned large language models can provide clinically meaningful, albeit imperfect, dermatology-related assistance [51,52]. These findings complement evaluations of fine-tuned models and support the broader claim that integrating vision-specific components—such as the sequential backbone investigated here—could strengthen emerging multimodal clinical workflows.

5. Generalization to a Different Medical Domain: The Kvasir Dataset

To further assess the robustness and versatility of the proposed SST framework, we evaluated its generalization to a different medical imaging domain. Among the few datasets providing both segmentation masks and classification labels, the Kvasir dataset [13] is particularly suitable due to its well-structured composition and clinical relevance in gastrointestinal endoscopy.

5.1. Kvasir Dataset Overview

Kvasir is a publicly available benchmark widely used for gastrointestinal image analysis, supporting both classification and segmentation. It comprises two complementary subsets:

- **Kvasir-Classification (KvasirC)** [13]: 8000 endoscopic images evenly distributed across eight diagnostic categories.

- **Kvasir-Segmentation (KvasirS)** [34]: 1000 high-resolution images with binary polyp masks.

Figure 10 shows representative KvasirC samples, including Dyed Lifted Polyps, Dyed Resection Margins, Esophagitis, Normal Cecum, Normal Pylorus, Normal Z-line, Polyps, and Ulcerative Colitis. The balanced class distribution and high-quality images make KvasirC well-suited for multiclass classification, while KvasirS focuses on polyp segmentation—crucial for colorectal cancer screening.

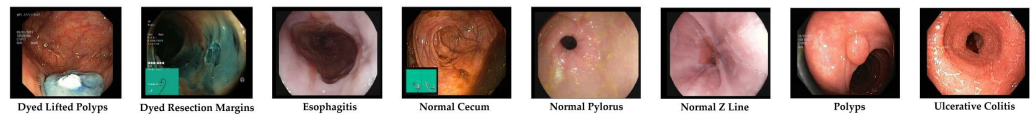


Figure 10. Representative gastrointestinal images from the Kvasir dataset, each annotated with its corresponding class label.

The two subsets are independent and do not share any overlapping images. We evaluated SST separately on KvasirC and KvasirS, following the same sequential setup as in HAM, to verify whether the benefits of task ordering persist across different imaging modalities. Both datasets were divided into 75% training, 15% validation, and 10% testing. Stratified sampling preserved class balance in KvasirC, and random splitting was applied to KvasirS. In the absence of an external test set, reproducibility was ensured by fixing random seeds.

5.2. Experimental Setup on Kvasir

We adopted the same STL-based training strategy used for HAM (Section 3), testing two task-order configurations:

- **Our_E:** Classification followed by segmentation.
- **Our_F:** Segmentation followed by classification.

Each configuration was trained independently, using the best validation checkpoint from the first task to initialize the second. Model architecture and training hyperparameters were kept identical to those optimized on the dermatology domain (HAM), without any further tuning on Kvasir, to ensure that observed gains were not driven by dataset-specific adjustments. Dataset partitioning for Kvasir followed the subject-level splits adopted in the previous literature to prevent data leakage across patients. All models were trained for 100 epochs per task, with an input size of 224×224 , a learning rate of 1×10^{-4} , and a batch size of 144. Results were averaged over five runs to ensure statistical reliability.

5.3. Quantitative Evaluation and Analysis

Table 6 reports the results on Kvasir, while Figure 11 shows confusion matrices and ROC curves. Evaluation metrics include Jaccard and Dice scores for segmentation, as well as accuracy, precision, recall, F1 score, and AUC for classification.

Table 6. Overall benchmark results of the SST model on the Kvasir datasets. TA denotes test accuracy. Jaccard and Dice scores evaluate segmentation performance, while TAb and TAm indicate binary and multiclass classification accuracy. TPm, TRm, and TF1m represent multiclass precision, recall, and F1 score.

Author	Jaccard (%)	Dice (%)	TA _b (%)	TA _m (%)	TP _m (%)	TR _m (%)	TF1 _m (%)
Our_E	90.95 ± 0.21	93.12 ± 0.39	99.4 ± 0.14	94.41 ± 0.37	94.54 ± 0.38	94.40 ± 0.38	94.38 ± 0.38
Our_F	91.12 ± 0.45	93.60 ± 0.17	98.9 ± 0.14	94.59 ± 0.41	94.16 ± 0.59	94.05 ± 0.60	94.03 ± 0.60

Segmentation: SST_SC achieved the best segmentation results, with 91.12% Jaccard and 93.60% Dice, slightly surpassing SST_CS (90.95% and 93.12%). Although the performance difference is modest, the consistency across runs suggests that learning segmentation

first improves the model's ability to capture fine-grained spatial structures such as polyp boundaries. This supports the hypothesis that early spatial learning benefits downstream feature extraction and classification stability.

Classification: Both configurations delivered strong classification results, with SST_SC achieving 94.59% multiclass accuracy and SST_CS 94.41%. Despite the small gap, the segmentation-first setup again shows an advantage, reinforcing the trend observed in HAM. ROC analyses (Figure 11) confirm these results, with both models yielding macro- and per-class AUCs above 0.97, and SST_SC displaying slightly improved discrimination in harder classes. These findings demonstrate that task ordering continues to influence representational focus, favoring spatial alignment and feature consistency, even in a different imaging context.

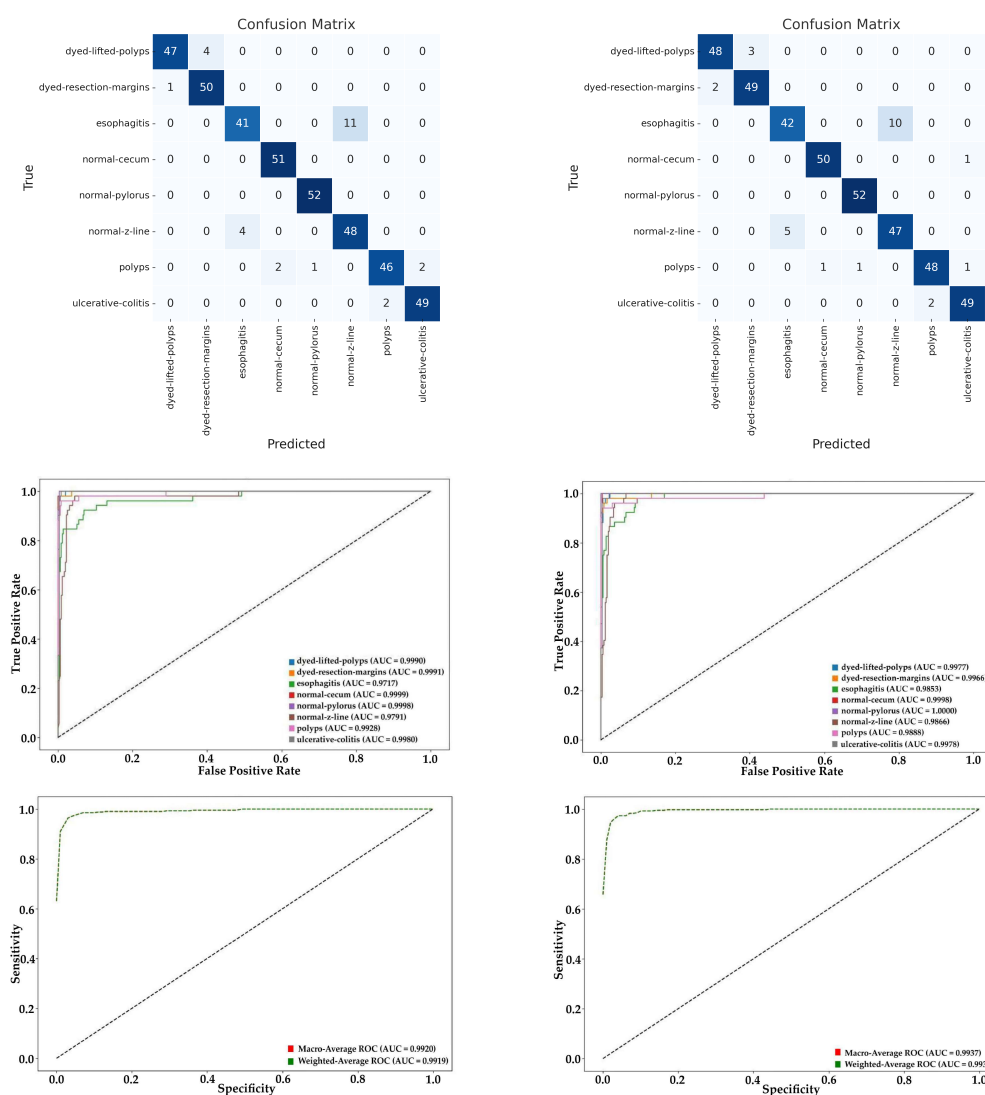


Figure 11. Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (from top to bottom) for the models Our_E (left column) and Our_F (right column), as reported in Table 6. Each row corresponds to the respective experimental setup, based on the best-performing checkpoint on the validation set. In the confusion matrices (top row), true labels are shown on the vertical axis and predicted labels on the horizontal axis; classes are indexed from 0 to 6, corresponding to MEL, NV, BCC, AKIEC, BKL, DF, and VASC, respectively. The ROC curves in the middle row display the True Positive Rate (TPR, sensitivity) on the y-axis versus the False Positive Rate (FPR) on the x-axis for each class. In the bottom row, macro and weighted average ROC curves are reported with the y-axis representing sensitivity and the x-axis representing specificity.

5.4. Comparison with Literature

To evaluate the competitiveness of SST beyond dermatology, we compared its performance with recent SOTA methods.

Segmentation: As reported in Table 7, SST_SC achieved the highest Jaccard (91.12%) and competitive Dice (93.60%), outperforming DUCK-net (90.51%) and all EffiSegNet variants [36]. While differences are limited to 0.5–1 percentage points, they are consistent and statistically stable, confirming the advantage of introducing segmentation first. These results highlight SST’s ability to reach top-tier performance despite not being explicitly tailored for this task.

Table 7. Segmentation benchmark on the KvasirS dataset. Evaluation based on Jaccard and Dice metrics.

Method	Jaccard (%)	Dice (%)
DUCK-net [35]	90.51	95.02
EffiSegNet-B4 [36]	90.56	94.83
EffiSegNet-B6 [36]	90.60	94.77
EffiSegNet-B5 [36]	90.65	94.88
SST_CS (Our_E)	90.95 ± 0.21	93.12 ± 0.39
SST_SC (Our_F)	91.12 ± 0.45	93.60 ± 0.17

Classification: Table 8 shows that SST_SC reached 94.59% accuracy, slightly surpassing SST_CS (94.41%) and outperforming several recent methods by Fonolla et al. [53] (90.20%), Zhang et al. [54] (90.40%), and Demirbas et al. [37] (93.37%). Although these margins are narrow, they consistently favor the segmentation-first approach, suggesting that integrating spatial priors early in training improves discriminative representation and domain robustness.

Table 8. Multiclass classification accuracy on the KvasirC dataset. TA_multiclass (%) indicates classification accuracy.

Method	TA_multiclass (%)
Multi-model classification [53]	90.20
Single Shot MultiBox Detector [54]	90.40
Transfer Learning framework [55]	93.00
Deep CNN-based SAM [56]	93.19
Spatial-attention ConvMixer [37]	93.37
SST_CS (Our_E)	94.41 ± 0.37
SST_SC (Our_F)	94.59 ± 0.41

6. Conclusions

This study introduced a sequential learning framework that systematically examines how the order of segmentation and classification affects representation learning in medical image analysis. Unlike traditional multitask approaches that optimize tasks jointly, the proposed method isolates task ordering within a shared Transformer backbone, providing a controlled setting to analyze inter-task feature transfer. Across experiments in dermatology (HAM) and gastrointestinal imaging (Kvasir), the configuration in which segmentation precedes classification consistently achieved higher accuracy, stronger cross-domain generalization, and more interpretable activations.

Quantitative metrics confirmed its advantage in both accuracy and F1 score, while Grad-CAM and t-distributed stochastic neighbor embedding (t-SNE) analyses revealed lesion-centered attention and well-separated feature clusters, indicating the learning of pathology-relevant representations. Compared to traditional preprocessing-based pipelines,

the proposed sequential strategy efficiently reuses shared parameters, maintains robust external performance, and emphasizes the importance of standardized and reproducible evaluation protocols in medical imaging.

From a theoretical perspective, this framework establishes a reproducible paradigm for studying inter-task transfer in deep networks, bridging the gap between single-task and multitask learning. Clinically, its ability to generate localized and transparent predictions enhances interpretability and diagnostic accountability. The current framework is not intended as a deployment-ready system. Key components needed for clinical deployment—such as probability calibration, uncertainty estimation, or gating, and resource-efficient variants of the backbone (e.g., distilled or quantized models)—were beyond the scope of this study and remain important directions for future work.

Current limitations primarily concern the partial alignment between segmentation and classification datasets, which constrains the full exploitation of shared spatial priors. Future work will focus on extending dataset coverage, refining task-specific heads, and exploring self-supervised or multimodal extensions to further improve generalization under data scarcity.

In summary, the proposed sequential approach offers a principled and interpretable route toward transferable and clinically reliable medical AI, demonstrating that segmentation-first learning fosters both stronger generalization and enhanced interpretability in deep visual analysis.

Author Contributions: M.G., I.G. and S.C. contributed equally to the conceptualization, methodology, software implementation, data analysis, and manuscript preparation. Supervision was jointly carried out by I.G. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets employed in this study are publicly available from their original sources, as cited in the manuscript (HAM10000 [5], and Kvasir [13]). All experiments were conducted in a containerized environment equipped with a single Nvidia A100 80GB GPU, 16 AMD EPYC 7742 CPU cores, 64 GB of RAM, and a 76 TB RAID 6 storage system. The complete implementation of the SST framework, including training scripts, model configurations, and the official split files for the HAM and Kvasir datasets, is now publicly available on [GitHub](#). Fixed random seeds were used to ensure full reproducibility. These resources enable direct replication and promote fair benchmarking across future studies.

Acknowledgments: During the preparation of this manuscript, the author(s) used “Grammarly” for the purposes of “language polishing”. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Waseh, S.; Lee, J.B. Advances in melanoma: Epidemiology, diagnosis, and prognosis. *Front. Med.* **2023**, *10*, 1268479. [[CrossRef](#)]
2. Woo, Y.R.; Cho, S.H.; Lee, J.D.; Kim, H.S. The human microbiota and skin cancer. *Int. J. Mol. Sci.* **2022**, *23*, 1813. [[CrossRef](#)]
3. Schadendorf, D.; Van Akkooi, A.C.; Berking, C.; Griewank, K.G.; Gutzmer, R.; Hauschild, A.; Stang, A.; Roesch, A.; Ugurel, S. Melanoma. *Lancet* **2018**, *392*, 971–984. [[CrossRef](#)] [[PubMed](#)]
4. Gallazzi, M.; Biavaschi, S.; Bulgheroni, A.; Gatti, T.; Corchs, S.; Gallo, I. A Large Dataset to Enhance Skin Cancer Classification with Transformer-Based Deep Neural Networks. *IEEE Access* **2024**, *12*, 109544–109559. [[CrossRef](#)]
5. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [[CrossRef](#)] [[PubMed](#)]

6. Gallazzi, M.; Rehman, A.U.; Corchs, S.; Gallo, I. Improving Classification in Skin Lesion Analysis Through Segmentation. In Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods, Beijing, China, 24–26 October 2025; pp. 696–703.
7. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med Imaging* **2016**, *35*, 1299–1312. [\[CrossRef\]](#)
8. He, X.; Wang, Y.; Zhao, S.; Chen, X. Joint segmentation and classification of skin lesions via a multi-task learning convolutional neural network. *Expert Syst. Appl.* **2023**, *230*, 120174. [\[CrossRef\]](#)
9. Mustafa, S.; Jaffar, A.; Rashid, M.; Akram, S.; Bhatti, S.M. Deep learning-based skin lesion analysis using hybrid ResUNet++ and modified AlexNet-Random Forest for enhanced segmentation and classification. *PLoS ONE* **2025**, *20*, e0315120. [\[CrossRef\]](#)
10. Khan, M.A.; Sharif, M.; Akram, T.; Damaševičius, R.; Maskeliūnas, R. Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics* **2021**, *11*, 811. [\[CrossRef\]](#)
11. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
12. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.
13. Pogorelov, K.; Randel, K.R.; Griwodz, C.; Eskeland, S.L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D.T.; Lux, M.; Schmidt, P.T.; et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan, 20–23 June 2017; pp. 164–169.
14. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
15. Masood, H.; Naseer, A.; Saeed, M. Optimized Skin Lesion Segmentation: Analysing DeepLabV3+ and ASSP Against Generative AI-Based Deep Learning Approach. *Found. Sci.* **2024**, *30*, 447–471. [\[CrossRef\]](#)
16. Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 699–711. [\[CrossRef\]](#)
17. Wu, H.; Chen, S.; Chen, G.; Wang, W.; Lei, B.; Wen, Z. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal.* **2022**, *76*, 102327. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Wang, J.; Wei, L.; Wang, L.; Zhou, Q.; Zhu, L.; Qin, J. Boundary-aware transformers for skin lesion segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Proceedings, Part I 24; Springer: Berlin/Heidelberg, Germany, 2021; pp. 206–216.
19. Hu, M.; Li, Y.; Yang, X. Skinsam: Empowering skin cancer segmentation with segment anything model. *arXiv* **2023**, arXiv:2304.13973. [\[CrossRef\]](#)
20. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [\[CrossRef\]](#)
21. Ali, M.S.; Miah, M.S.; Haque, J.; Rahman, M.M.; Islam, M.K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **2021**, *5*, 100036. [\[CrossRef\]](#)
22. Datta, S.K.; Shaikh, M.A.; Srihari, S.N.; Gao, M. Soft attention improves skin cancer classification performance. In Proceedings of the Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, 27 September 2021; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2021; pp. 13–23.
23. Ayas, S. Multiclass skin lesion classification in dermoscopic images using swin transformer model. *Neural Comput. Appl.* **2023**, *35*, 6713–6722. [\[CrossRef\]](#)
24. Xin, C.; Liu, Z.; Zhao, K.; Miao, L.; Ma, Y.; Zhu, X.; Zhou, Q.; Wang, S.; Li, L.; Yang, F.; et al. An improved transformer network for skin cancer classification. *Comput. Biol. Med.* **2022**, *149*, 105939. [\[CrossRef\]](#)
25. Combalia, M.; Codella, N.C.; Rotemberg, V.; Helba, B.; Vilaplana, V.; Reiter, O.; Carrera, C.; Barreiro, A.; Halpern, A.C.; Puig, S.; et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv* **2019**, arXiv:1908.02288. [\[CrossRef\]](#)
26. Lan, Z.; Cai, S.; He, X.; Wen, X. FixCaps: An improved capsules network for diagnosis of skin cancer. *IEEE Access* **2022**, *10*, 76261–76267. [\[CrossRef\]](#)
27. Gururaj, H.L.; Manju, N.; Nagarjun, A.; Aradhya, V.M.; Flammini, F. DeepSkin: A deep learning approach for skin cancer classification. *IEEE Access* **2023**, *11*, 50205–50214. [\[CrossRef\]](#)

28. Yang, J.; Wang, J.; Zhang, G.; Li, Y. Selecting the best sequential transfer path for medical image segmentation with limited labeled data. *arXiv* **2024**, arXiv:2410.06892. [[CrossRef](#)]
29. Akram, A.; Rashid, J.; Jaffar, M.A.; Faheem, M.; Amin, R.U. Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things. *Skin Res. Technol.* **2023**, *29*, e13524. [[CrossRef](#)] [[PubMed](#)]
30. Wang, Y.; Su, J.; Xu, Q.; Zhong, Y. A collaborative learning model for skin lesion segmentation and classification. *Diagnostics* **2023**, *13*, 912. [[CrossRef](#)] [[PubMed](#)]
31. Chowdary, J.; Yogarajah, P.; Chaurasia, P.; Guruviah, V. A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images. *Ultrason. Imaging* **2022**, *44*, 3–12. [[CrossRef](#)]
32. Manzoor, K.; Gilal, N.U.; Agus, M.; Schneider, J. Dual-stage segmentation and classification framework for skin lesion analysis using deep neural network. *Digit. Health* **2025**, *11*, 20552076251351858. [[CrossRef](#)] [[PubMed](#)]
33. Himel, G.M.S.; Islam, M.M.; Al-Aff, K.A.; Karim, S.I.; Sikder, M.K.U. Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system. *Int. J. Biomed. Imaging* **2024**, *2024*, 3022192. [[CrossRef](#)]
34. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; De Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the MultiMedia modeling: 26th International Conference, MMM 2020, Daejeon, Republic of Korea, 5–8 January 2020; Proceedings, Part II 26; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
35. Dumitru, R.G.; Peteleaza, D.; Craciun, C. Using DUCK-Net for polyp image segmentation. *Sci. Rep.* **2023**, *13*, 9803. [[CrossRef](#)]
36. Vezakis, I.A.; Georgas, K.; Fotiadis, D.; Matsopoulos, G.K. EffiSegNet: Gastrointestinal Polyp Segmentation through a Pre-Trained EfficientNet-based Network with a Simplified Decoder. *arXiv* **2024**, arXiv:2407.16298.
37. Demirbaş, A.A.; Üzen, H.; Firat, H. Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset. *Health Inf. Sci. Syst.* **2024**, *12*, 32. [[CrossRef](#)]
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
39. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368. [[CrossRef](#)]
40. 2018, I.C. HAM10000 Test Set Release. Available online: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T> (accessed on 20 May 2025).
41. Benčević, M.; Galić, I.; Habijan, M.; Babin, D. Training on polar image transformations improves biomedical image segmentation. *IEEE Access* **2021**, *9*, 133365–133375. [[CrossRef](#)]
42. Aladhadh, S.; Alsanea, M.; Aloraini, M.; Khan, T.; Habib, S.; Islam, M. An effective skin cancer classification mechanism via medical vision transformer. *Sensors* **2022**, *22*, 4008. [[CrossRef](#)] [[PubMed](#)]
43. Mushtaq, S.; Singh, O. A deep learning based architecture for multi-class skin cancer classification. *Multimed. Tools Appl.* **2024**, *83*, 87105–87127. [[CrossRef](#)]
44. Jain, S.; Singhanian, U.; Tripathy, B.; Nasr, E.A.; Aboudaif, M.K.; Kamrani, A.K. Deep learning-based transfer learning for classification of skin cancer. *Sensors* **2021**, *21*, 8142. [[CrossRef](#)]
45. Chaturvedi, S.S.; Tembhurne, J.V.; Diwan, T. A multi-class skin Cancer classification using deep convolutional neural networks. *Multimed. Tools Appl.* **2020**, *79*, 28477–28498.
46. Shetty, B.; Fernandes, R.; Rodrigues, A.P.; Chengoden, R.; Bhattacharya, S.; Lakshmana, K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci. Rep.* **2022**, *12*, 18134. [[PubMed](#)]
47. ISIC Collaboration (2016–2024). ISIC Challenge Datasets. Available online: <https://challenge.isic-archive.com/data/> (accessed on 20 November 2025).
48. Yu, X.; Liang, X.; Zhou, Z.; Zhang, B. Multi-task learning for hand heat trace time estimation and identity recognition. *Expert Syst. Appl.* **2024**, *255*, 124551. [[CrossRef](#)]
49. Zhou, J.; He, X.; Sun, L.; Xu, J.; Chen, X.; Chu, Y.; Zhou, L.; Liao, X.; Zhang, B.; Gao, X. SkinGPT-4: An interactive dermatology diagnostic system with visual large language model. *arXiv* **2023**, arXiv:2304.10691. [[CrossRef](#)]
50. Zhou, J.; He, X.; Sun, L.; Xu, J.; Chen, X.; Chu, Y.; Zhou, L.; Liao, X.; Zhang, B.; Afvari, S.; et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **2024**, *15*, 5649. [[CrossRef](#)]
51. Boostani, M.; Bánvölgyi, A.; Zouboulis, C.C.; Goldfarb, N.; Suppa, M.; Goldust, M.; Lőrincz, K.; Kiss, T.; Nádudvari, N.; Holló, P.; et al. Large language models in evaluating hidradenitis suppurativa from clinical images. *J. Eur. Acad. Dermatol. Venereol. JEADV* **2025**, *39*, e1052–e1055.
52. Boostani, M.; Bánvölgyi, A.; Goldust, M.; Cantisani, C.; Pietkiewicz, P.; Lőrincz, K.; Holló, P.; Wikonkál, N.M.; Paragh, G.; Kiss, N.; et al. Diagnostic Performance of GPT-4o and Gemini Flash 2.0 in Acne and Rosacea. *Int. J. Dermatol.* **2025**, *64*, 1881–1882. [[CrossRef](#)] [[PubMed](#)]

53. Fonolla, R.; van der Sommen, F.; Schreuder, R.M.; Schoon, E.J.; de With, P.H. Multi-modal classification of polyp malignancy using CNN features with balanced class augmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, Venice, Italy, 8–11 April 2019; pp. 74–78.
54. Zhang, X.; Chen, F.; Yu, T.; An, J.; Huang, Z.; Liu, J.; Hu, W.; Wang, L.; Duan, H.; Si, J. Real-time gastric polyp detection using convolutional neural networks. *PLoS ONE* **2019**, *14*, e0214133. [[CrossRef](#)]
55. Liu, X.; Wang, C.; Bai, J.; Liao, G. Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images. *Neurocomputing* **2020**, *392*, 253–267. [[CrossRef](#)]
56. Lonseko, Z.M.; Adjei, P.E.; Du, W.; Luo, C.; Hu, D.; Zhu, L.; Gan, T.; Rao, N. Gastrointestinal disease classification in endoscopic images using attention-guided convolutional neural networks. *Appl. Sci.* **2021**, *11*, 11136. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.