

Improving Classification in Skin Lesion Analysis Through Segmentation

Mirco Gallazzi^a, Anwar Ur Rehman^b, Silvia Corchs^c and Ignazio Gallo^d

Department of Theoretical and Applied Science, University of Insubria, 21100 Varese, Italy
{mgallazzi2, aurehman1, silvia.corchs, ignazio.gallo}@uninsubria.it

Keywords: YOLO, Swin Transformer, Object Detection and Segmentation, Medical Imaging, Dermatology, Skin Cancer.

Abstract: Deep Learning plays a vital role in medical imaging, especially in classification and segmentation tasks essential for diagnosing diseases from images. However, current methods often struggle to differentiate visually similar classes and accurately delineate lesion boundaries. This study builds on prior findings of classification limitations, investigating whether segmentation can improve classification performance for skin lesion analysis with Transformer-based models. We benchmarked the segmentation capabilities of the Swin Transformer, YOLOv8, and DeepLabV3 architectures on the HAM dataset, which contains 10,015 images across seven skin lesion classes. Swin outperformed others in segmentation, achieving an intersection over union of 82.75%, while YOLOv8 achieved 77.0%. However, classification experiments using classification datasets after segmenting and cropping the lesion of interest did not produce the expected improvements, with classification accuracy showing slight drops in the segmented data. For example, on the original HAM dataset, the model achieved a Test Accuracy (TA) of 84.64%, while Swin trained on segmented data showed a slight decline to a TA of 84.13%. These findings suggest that segmentation alone may not effectively support classification. Based on this, we propose future research into a sequential transfer learning approach, where segmentation knowledge could be progressively transferred to improve classification.

1 INTRODUCTION


In medical imaging, classification and segmentation support in-depth disease analysis. Classification categorizes images by pathology, while segmentation highlights specific structures like lesions, adding spatial information that may improve class differentiation in dermatology.


Advances in Deep Learning (DL), such as those demonstrated by Esteva et al. (Esteva et al., 2017), have underscored both the potential and limitations of classification in skin lesion analysis, particularly in managing nuanced visual distinctions essential for accurate diagnosis. It has been shown in the literature that Transformer-based models, specifically the Swin Transformer (Swin) (Liu et al., 2021), achieved high classification accuracy across multiple skin lesion types, affirming the model's promise in skin cancer analysis (Gallazzi et al., 2024). However, persistent challenges remain, particularly in differentiating between classes with subtle visual similarities, such


as Nevus (NV) and Melanoma (MEL), suggesting that classification alone may lack the precision required for reliable distinction.


Given these limitations, we hypothesize that segmentation, when applied as a supportive task, could enhance classification performance by embedding spatial context around lesion boundaries and structures, providing a refined feature set for complex lesion types. This study begins by investigating segmentation as an isolated task, examining its effectiveness on the HAM dataset (Tschandl and Rosendahl, 2018), which comprises over 10,000 images across seven skin lesion classes. We benchmark the segmentation capabilities of Swin, YOLOv8 (Ultralytics, 2024), and a ResNet(DeepLabV3) (He et al., 2016), assessing whether segmentation independently contributes meaningful improvements. Following this evaluation, we use cropped images of the segmented regions from each model as input for classification tasks, enabling a direct comparison of classification accuracy with and without segmented data.

Our findings reveal that using cropped images from segmented regions alone does not significantly enhance classification performance and may even introduce minor declines, suggesting that segmentation

^a  <https://orcid.org/0009-0000-3850-8086>

^b  <https://orcid.org/0000-0002-9384-8988>

^c  <https://orcid.org/0000-0002-1739-8110>

^d  <https://orcid.org/0000-0002-7076-8328>

and classification independently may not sufficiently enhance model accuracy. These observations lead us to consider a Sequential Transfer Learning (STL) approach, where the knowledge from segmentation of cropped regions is sequentially transferred to classification. This framework could allow segmentation to embed boundary and structural details that classification can then refine, addressing the interpretative challenges presented by visually similar lesion types. This approach raises two primary questions: *“How can segmentation knowledge be effectively transferred to improve classification performance?”* and *“In what ways does a sequential integration of segmentation and classification impact interpretative accuracy and robustness in skin lesion analysis?”*

The main contributions are:

1. Compare Swin’s segmentation performance relative to YOLOv8 and DeepLabV3, assessing its utility in medical imaging.
2. Evaluating whether using cropped images derived from segmentation enhances class separability when used as a preprocessing step for classification.
3. Discuss the potential of an STL-based integration, discussing how findings from our study and relevant literature support task sequence alignment for performance gains in future work.

Our paper is structured as follows: Section 2 reviews current approaches in segmentation tasks. Section 3 provides details on the HAM dataset and model architectures, and Section 4 presents our experimental results. Finally, Section 5 discusses our findings and outlines future directions for a potential STL pipeline in skin lesion detection.

2 RELATED WORKS

DL models have shown significant success in both classification and segmentation tasks, enhancing performance in various real-time applications such as autonomous driving, agriculture (Gallo et al., 2023), industrial automation (Rehman and Gallo, 2024)(Rehman et al., 2023), and especially medical imaging. Dermatology, in particular, benefits from these advancements, yet certain nuances in skin lesion differentiation remain challenging, necessitating further exploration of methods that could improve class separability in complex cases(Gallazzi et al., 2024).

Convolutional Neural Networks (CNNs) (O’Shea, 2015) and Transformer-based architectures have yielded strong results in skin lesion classification,

often using datasets like HAM. Esteva et al. (Esteva et al., 2017) developed a CNN for skin cancer classification with performance comparable to dermatologists, marking a significant milestone in the field. ResNet, introduced in (He et al., 2016), is a widely adopted CNN architecture in medical imaging due to its residual connections, which allow for deeper networks and improved detection of subtle differences. Integrated into encoder-decoder frameworks, such as U-Net (Ronneberger et al., 2015), DeepLabv3’s deep layers enable robust feature extraction and fine-grained segmentation.

U-Net’s encoder-decoder architecture with skip connections has become foundational in medical segmentation due to its precise spatial reconstruction capabilities. Adaptations of U-Net with attention mechanisms and multi-scale processing (Azad et al., 2024) have addressed challenges specific to complex medical images, including dermatology.

The adaptation of Transformer-based models, particularly Swin, has advanced segmentation accuracy by capturing global dependencies within high-resolution images. Liu et al. (Liu et al., 2021) demonstrated Swin’s strengths in managing complex lesion boundaries in dermoscopic images. Swin-Unet (Wang et al., 2022) combines Swin’s attention features with U-Net’s localization capabilities, making it highly suitable for anatomically complex regions in challenging datasets.

YOLO-based models, originally developed for object detection, have been adapted to skin lesion segmentation to address scale and feature extraction challenges. With real-time processing capabilities and the integration of attention mechanisms, YOLO models demonstrate increased sensitivity to lesion boundaries, facilitating effective segmentation in high-resolution dermoscopic images. Multi-step approaches, such as the combination of YOLO and SegNet proposed in (Taghizadeh and Mohammadi, 2022), have been shown to optimize segmentation accuracy.

Given these advancements, the potential of STL has become an area of interest, especially for tasks where segmentation might support or refine classification. STL involves training a model on one task, such as segmentation, to capture intricate structural features before fine-tuning it for classification. This approach could leverage the strengths of each task sequentially. According to (Mao, 2020), this framework may allow models to better adapt knowledge from segmentation to classification, addressing class separability challenges in visually complex datasets.

3 METHODOLOGY

Following the Introduction 1 and Related Works 2, we now delve into the challenges of applying STL to skin cancer classification, focusing on distinguishing between visually similar classes. Although transformers like Swin have shown high accuracy across multiple medical imaging tasks, a detailed T-Distributed Stochastic Neighbor Embedding (t-SNE) (Cai and Ma, 2022) visualization of the feature space reveals an inherent limitation in this context in Figure 1. The plot shows that lesion classes with subtle visual similarities—such as Nevus and Melanoma—tend to cluster closely, leading to substantial overlap. This clustering highlights the challenge of forming well-defined boundaries between classes that share intricate visual patterns, underscoring a critical gap in model interpretability and reliability.

This limitation prompts an investigation into whether segmentation can enhance class separability by introducing spatial details and isolating key characteristics that classification struggles to distinguish. Segmentation, by improving boundary detection and emphasizing lesion-specific features, could provide a refined representation of each class, potentially enhancing classification outcomes.

In the following subsections, we examine the dataset and model architectures used in this study to address the challenges. Specifically, we explore the effectiveness of YOLO, Swin, and DeepLabV3 models—each selected for its unique advantages in segmentation and/or classification. YOLO’s strengths in real-time object detection make it a valuable candidate for efficient segmentation, particularly in high-resolution dermoscopic images, bringing great benefits to creating a segmented dataset through its use. With a slight modification to the original architecture, Swin could effectively capture long-range dependencies by handling complex boundary delineations within high-resolution images. DeepLabV3, with its residual connections, is widely adopted in medical imaging and serves as a robust baseline for classification and segmentation tasks due to its strong feature extraction capabilities. Together, these models cover a range of learning frameworks (hierarchical, residual, and attention-based), offering a comprehensive approach for testing segmentation’s impact on classification. Section 2 has highlighted the adaptability of these architectures across various medical imaging tasks, and here, we aim to assess their performance specifically on the HAM dataset, which provides both classification labels and segmentation masks to support a comprehensive evaluation across both tasks.

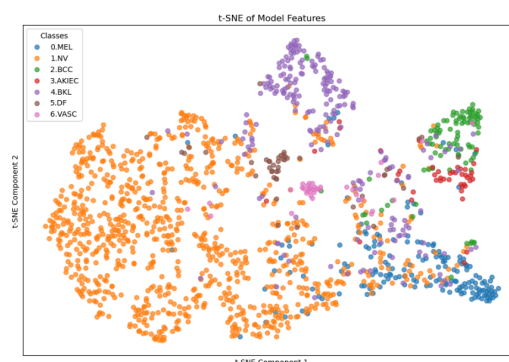


Figure 1: T-SNE visualization of class embeddings, illustrating the significant overlap between visually similar skin lesion classes, underscoring the challenge of achieving clear class separability in classification tasks.

3.1 Human Against Machine (HAM)

The HAM dataset, which includes both classification labels and segmentation masks across seven diverse dermatological classes, serves as a benchmark in our study, allowing a robust evaluation of model performance in analyzing distinct yet visually similar lesion types.

The training dataset encompasses 10,015 images distributed across the following seven classes: MEL, NV, Basal Cell Carcinoma (BCC), Actinic Keratoses (AKIEC), Benign Keratosis Like (BKL), Dermatofibroma (DF), Vascular (VASC) comprises 1,113, 6,705, 514, 327, 1,099, 115, and 142 images respectively. Each class represents a unique diagnostic category, contributing to the dataset’s complexity and offering a suitable challenge for classification and segmentation models. Figure 2 illustrates an example image from each class.

HAM also provides two distinct external test sets for classification and segmentation to enable a thorough evaluation. The classification test set comprises 1,511 images, categorized as follows: 171 MEL, 908 NV, 93 BCC, 43 AKIEC, 217 BKL, 44 DF, and 35 VASC. This distribution reflects the variety and imbalance in clinical datasets, testing model robustness across less frequent classes and those with visual similarities.

HAM provides a separate segmentation test set of 1,000 images, each with a corresponding segmentation mask. This segmentation set is critical for assessing model capabilities in identifying and isolating lesion boundaries. We hypothesize that this task may help address classification limitations by providing spatially precise lesion delineations.

By utilizing both test sets, we evaluate our models on classification and segmentation performance, comparing how each approach handles the specific chal-

lenges posed by the HAM dataset's complex and diverse lesion types.

3.2 Deep Models: DeepLabV3 vs Swin

DeepLabV3 and Swin are adapted into encoder-decoder structures for effective pixel-wise predictions in segmentation tasks. DeepLabV3's convolutional backbone is the encoder, extracting increasingly abstract features through residual blocks, while the decoder restores spatial resolution via transposed convolutions or upsampling layers. Skip connections retain low-level details, enhancing boundary accuracy. The residual function in each encoder block processes the input X as:

$$F(X) = \sigma(WX + b) + X \quad (1)$$

where $F(X)$ is the output combining learned transformations and original input, W and b are the weight matrix and bias, and σ is the non-linear activation function.

This formulation allows gradients to bypass the transformation layer, ensuring stable training even in deep networks, as the residual connections mitigate vanishing gradient issues.

In contrast, the Swin adapts its hierarchical transformer blocks into an encoder-decoder framework specifically suited for segmentation by utilizing window-based multi-head self-attention within the encoder. This segmentation approach applies self-attention within small, shifted windows, capturing spatial dependencies locally while maintaining computational efficiency. The self-attention computation for each window can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

where Q , K , and V represent the query, key, and value matrices, and d is the scaling factor. Swin's decoder restores spatial resolution with patch embeddings, blending local details and global context. Its multi-scale architecture aligns with the encoder-decoder paradigm, enhancing boundary delineation at multiple resolutions.

For the decoder, Swin applies patch embedding layers to restore the spatial resolution, using attention mechanisms to integrate local detail with global context. Furthermore, Swin's multi-scale structure provides a natural fit for the encoder-decoder paradigm, with feature maps at varying resolutions feeding into corresponding levels of the decoder to produce segmentation maps with accurate boundary delineations.

Given that self-attention in Swin's encoder scales with complexity:

$$O((H \times W) \times d^2) \quad (3)$$

where H and W denote the height and width of the input feature map, and d the window size.

This complexity highlights the balance Swin achieves between global context capture and computational feasibility, using a manageable number of attention computations within each window.

ResNet's convolutional encoder is proficient at identifying localized patterns, ideal for tasks with clear boundaries. Swin, however, captures long-range dependencies and complex spatial relationships through attention mechanisms, excelling in images requiring both local and global context. While ResNet-based decoders often use convolutional upsampling, Swin's attention-based upsampling integrates multi-scale global features, improving boundary and texture precision in segmentation outputs.

3.3 You Only Look Once (YOLO)

YOLO is a family of models designed for real-time object detection that treats the detection task as a regression problem rather than a classification problem combined with region proposals (Redmon, 2016) (Rehman and Gallo, 2024). Unlike previous object detection methods, YOLO divides the image into an $S \times S$ grid, and each grid cell is responsible for detecting objects whose centers fall within it. For each grid cell, YOLO predicts B bounding boxes, a confidence score, and class probabilities. The key benefits of YOLO models are their speed and efficiency, allowing real-time detection even on limited hardware. This paper adopted the YOLOv8 model, which is the latest version of the YOLO series of object detection models, designed to enhance the accuracy and speed of the previous versions. YOLOv8 improves over YOLOv7 with more efficient feature extraction and better loss function optimization (Ultralytics, 2024). Unlike its predecessors, YOLOv8 can perform classification, object detection, and instance segmentation, making it an all-in-one solution for computer vision tasks.

YOLOv8 introduces several key improvements compared to earlier versions, including *an improved backbone network* for feature extraction, *an advanced anchor-free detection head*, *enhanced loss functions* for more accurate training, and *additional support for instance segmentation tasks*.

The YOLOv8 architecture is built upon three main components: the backbone, the neck, and the head.

The **backbone** extracts features from the input image. In YOLOv8, CSP (Cross Stage Partial) networks (Wang et al., 2020) is used as the backbone, which reduces computational complexity while maintaining feature quality. The input image of size $W \times H \times 3$ is

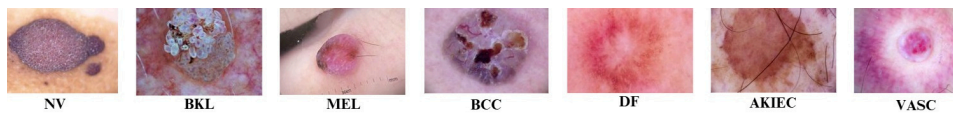


Figure 2: Examples of each of the seven classes of skin disease images from the HAM dataset.

transformed into a set of feature maps through convolutional layers and down-sampling operations.

The **neck** aggregates features from different backbone levels to improve detection accuracy for objects of different sizes. YOLOv8 employs a PANet (Path Aggregation Network) (Liu et al., 2018) that ensures efficient information flow from high-resolution to low-resolution layers and vice versa.

The detection **head** predicts bounding boxes, objectness scores, class labels, and segmentation masks. Unlike anchor-based approaches in earlier versions, YOLOv8 uses an anchor-free mechanism that directly predicts the object’s center, eliminating the complexity associated with anchor boxes.

Loss Function: YOLOv8’s training relies on several loss functions to optimize object detection and segmentation accuracy. However, the most commonly used loss function is GIoU loss.

It aims to provide a better measure for non-overlapping boxes by considering the smallest enclosing box C :

$$L_{GIoU} = 1 - IoU + \frac{|C - (B_p \cup B_g)|}{|C|}. \quad (4)$$

4 EXPERIMENTS AND RESULTS

Our primary goal in implementing Swin, DeepLabV3, and YOLO architectures on the HAM dataset was to explore whether using segmentation to derive cropped images could enhance classification outcomes, especially in dermatological settings where visually similar lesion classes are challenging to differentiate. By studying the segmentation and classification performance using the segmented images that these models generate, this investigation seeks to confirm the generalization capabilities and robustness of each model under the unique constraints and characteristics of HAM data.

To assess segmentation and classification performance on the HAM dataset, we employed several key metrics: Intersection over Union (IoU) for segmentation accuracy and accuracy, precision, recall, and F1-score for classification. Each metric offers specific insights, with IoU indicating the extent of overlap between predicted and ground-truth regions in segmentation and the other metrics assessing the accuracy and robustness of classification across classes with

varying visual similarities.

The metrics used are defined as follows:

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

$$Accuracy = \frac{TP + TN}{Total\ Samples} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

Where ”TP” indicates the ”True Positives”, ”TN” the ”True Negatives”, ”FP” ”FN” indicates the ”False Positives” and the ”False Negatives”, respectively.

All computations were performed on a containerized environment utilizing 1 Nvidia A100 80GB GPU, 16 cores of an AMD Epyc 7742 64-core CPU, and 64GB of DDR4-3200 RAM, all connected to a 76TB RAID6 storage server via a 25Gbps low-latency network. The environment used in our experiments uses PyTorch version 2.1.0 with CUDA version 11.8 and Python version 3.11.6.

4.1 Yolo vs ResNet/Swin in Segmentation

Table 1: Segmentation Results on HAM Test Set.

Experiment Type	Model	IoU (%)
Seg(Challenge)	MaskRcnn2	80.2
Seg	DeepLabV3	81.66
Seg	Swin	82.75
Seg	YOLO	77.0

In our first experiment, we evaluate the segmentation capabilities of three models—YOLOv8, DeepLabV3 architecture (Chen et al., 2018), and Swin—on the HAM dataset. The dataset was split into 80% for training and 20% for validation. Each model was trained to learn to segment skin lesions by directly comparing predictions against the true segmentation masks provided in HAM. After each training epoch, model performance was assessed on the external HAM segmentation test set, identical to the test set used in the HAM segmentation challenge (ISIC Challenge, 2024). This alignment enables a direct benchmark comparison, allowing us to assess our results in the context of the 2018 competition standards.

Table 2: Classification results on the HAM test set. TA, TP, TR, and TF1 indicate Test Accuracy, Test Precision, Test Recall, and Test F1 Score (%). The "Dataset" column shows the HAM dataset segmented by different models.

Experiment Type	Dataset	Model	TA (%)	TP (%)	TR (%)	TF1 (%)
Classification	HAM	Swin	84.64	84.77	84.64	84.57
Classification	HAM_segmented_YOLO	Swin	84.01	84.09	84.12	83.53
Classification	HAM_segmented_DeepLabV3	Swin	84.12	84.38	84.12	83.75
Classification	HAM_segmented_Swin	Swin	84.13	84.41	84.12	83.58

Segmentation performance was evaluated using the IoU (5) metric, consistent with the evaluation criteria used in the HAM challenge. In the challenge, the winning score reached an IoU of 80.2%. Our results showed in Table 1 indicate improved performance, with DeepLabV3 achieving an IoU of 81.66% and Swin reaching 82.75% while YOLO implementation didn't surpass the bench value, achieving a 77.0%. These results highlight residual and transformer-based architectures' effectiveness in medical image segmentation, especially in distinguishing complex lesion boundaries.

The segmentation experiment underscores the capacity of Swin and DeepLabV3-based architectures to generalize effectively in medical segmentation tasks, establishing a strong foundation for further integration of segmentation into a comprehensive classification task.

4.2 Yolo vs ResNet/Swin in Classification

Following the results of the segmentation experiment in 4.1, we created three separate versions of cropped HAM images, each derived from the segmentation masks generated by the three models: YOLO, DeepLabV3, and Swin. Each model produced a separate segmented dataset, leveraging its own segmentation capabilities to generate unique annotations for the HAM images. This approach allowed us to explore how different segmentation models might influence classification performance when used as preprocessed inputs.

Each of these newly created datasets of cropped images from the segmented regions was subsequently fed into the Swin model to train it on classification, aiming to assess any potential improvements in distinguishing between dermatological multi-classes. In this experiment, we evaluated classification performance using accuracy (6), precision (7), recall (8), and f1-score (9) to gauge whether segmentation could enhance the Swin model's ability to differentiate between classes with subtle visual similarities.

The results presented in Table 2 show that using cropped images derived from segmented regions of the HAM dataset did not yield a meaningful improve-

ment in classification performance. The Swin model achieved a Test Accuracy (TA) of 84.64%, Test Precision (TP) of 84.77%, Test Recall (TR) of 84.64%, and Test F1 Score (TF1) of 84.57% when trained on the original HAM dataset, serving as a benchmark.

In contrast, when trained on cropped images segmented by each of the three models—YOLO, DeepLabV3, and Swin itself—performance declined slightly across all metrics. The YOLO-segmented dataset yielded a TA of 84.01%, TP of 84.09%, TR of 84.12%, and TF1 of 83.53%. DeepLabV3's segmented dataset resulted in a TA of 84.12%, TP of 84.38%, TR of 84.12%, and TF1 of 83.75%. Similarly, Swin's own segmented dataset achieved a TA of 84.13%, TP of 84.41%, TR of 84.12%, and TF1 of 83.58%.

These findings indicate that using segmented data for classification may introduce noise or obscure critical details, particularly in classes with similar visual features. This leads to slightly reduced accuracy and other metrics. This outcome underscores the limitations of using segmented images alone for classification tasks. It reinforces the need for an approach that can sequentially optimize segmentation and classification within a unified framework. Moreover, while segmentation provides valuable spatial details, its role as a preprocessing step for classification did not yield the expected improvements. This study thus motivates a shift toward STL as a potential solution for sequentially combining these tasks to enhance overall model performance in future work.

4.3 Integrating Segmentation in Classification: Discussion

The experiments in Section 4.1 confirmed that the selected models, particularly Swin, perform effectively in skin lesion segmentation, demonstrating robustness in capturing nuanced boundaries and structural details important for medical imaging. This supports Swin's utility in segmentation tasks, where precise delineation is essential for reliable lesion analysis. The strong performance in segmentation highlights the potential of Swin to capture spatial details that could be valuable for other tasks, such as classifica-

However, as shown in Section 4.2, our subsequent classification experiments—using datasets segmented by each model—did not yield the anticipated improvements. This indicates that the segmented datasets may have removed critical information or introduced noise, which is especially problematic when distinguishing between visually similar lesion classes. These results highlight a potential limitation in using segmented images alone for classification, particularly when segmentation is treated as a preprocessing step rather than being integrated into a unified framework.

Given these findings, an intriguing question arises: If segmentation alone does not improve classification outcomes, could sequentially transferring knowledge from segmentation to classification enhance performance? Recent research on Sequential Transfer Learning (STL) supports this hypothesis. Studies such as Paulsen and Casey (Paulsen and Casey, 2023) have shown the benefits of STL, where segmentation pre-training improves the classification of complex visual classes by leveraging spatial information gained in the initial phase. Similarly, Chan et al. (Chan et al., 2023) demonstrated that STL approaches involving pre-training on large datasets before fine-tuning on target tasks can significantly enhance model accuracy. Tirinzoni et al. (Tirinzoni et al., 2020) further highlighted that spatial insights gained in the first stage of STL can be effectively transferred to improve performance in downstream tasks.

In addition to these findings, Wang et al. (Wang et al., 2023) proposed a Collaborative Learning Deep Convolutional Neural Networks model, which emphasizes the interdependence between segmentation and classification tasks. Their work demonstrates that segmentation can improve classification by providing lesion contour information, while classification can enhance segmentation through target localization maps. This highlights the potential of collaborative learning to exploit the correlation between tasks, particularly when sample data are limited. The findings from Wang et al. further motivate exploring methods like STL that sequentially or collaboratively integrate segmentation and classification.

Building on these insights, we propose an STL framework that first trains the Swin model on segmentation to capture essential boundary and structural characteristics, followed by fine-tuning for classification. Unlike traditional approaches that treat these tasks in isolation, STL allows the model to progressively refine its feature representations, leveraging segmentation-derived spatial insights to enhance lesion differentiation in classification.

Future work will focus on evaluating the impact of

this STL approach, aiming to establish a more accurate and robust pipeline for skin lesion analysis. The core idea is that STL can capitalize on the strengths of each task in sequence, with the hypothesis that Swin’s performance can be enhanced by incrementally refining features relevant to segmentation and classification, thereby capturing finer details that may be overlooked in a traditional isolated setup.

5 CONCLUSIONS

This study aimed to evaluate the Swin model’s effectiveness in skin lesion classification using segmented images, to benchmark Swin’s segmentation capabilities alongside YOLO and DeepLabV3 architectures, and to explore the potential for a segmentation-classification pipeline based on a sequential approach in dermatological analysis. The experiments provided valuable insights into each of these objectives.

Firstly, we assessed the standalone effectiveness of Swin, YOLO, and DeepLabV3 for skin lesion segmentation, comparing their performances on the HAM dataset. Our results suggest that Swin performs competitively with the other models, effectively managing complex boundary details and demonstrating its potential as a reliable tool for medical segmentation tasks.

Secondly, we investigated whether segmentation could improve class separability in classification tasks by using cropped images of the segmented regions as inputs for classification. The findings indicated that, while segmentation captures detailed structural information, using segmented data alone did not improve classification accuracy. This outcome suggests that segmentation, although beneficial for boundary delineation, may inadvertently remove essential contextual information needed for distinguishing between visually similar lesion classes.

This study introduces the potential of STL for sequentially aligning segmentation and classification. Using segmentation to first capture boundary details and then refine classification may improve model accuracy by incrementally enhancing task-specific features. This suggests a promising path for robust, unified models in skin cancer detection.

However, adopting an STL approach also introduces challenges. Sequentially transferring segmentation features to classification may amplify noise or propagate inaccuracies from the segmentation phase, especially if segmentation boundaries are imprecise. Furthermore, balancing the computational cost of sequential training and the risk of overfitting in each task stage are essential considerations that require

careful assessment. Addressing these challenges will be crucial to validate the effectiveness of STL and ensure that it generalizes well across various dermatological datasets.

Looking forward, we aim to investigate whether this STL approach can be generalized beyond medical imaging or whether its effectiveness is uniquely suited to clinical applications. Understanding its adaptability to broader contexts could reveal new possibilities for versatile models capable of handling complex visual recognition tasks across various domains.

REFERENCES

- Azad, R., Aghdam, E. K., Rauland, A., and Bozorgpour, A. (2024). Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Cai, T. T. and Ma, R. (2022). Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 23(301):1–54.
- Chan, J. Y.-L., Bea, K. T., Leow, S. M. H., Phoong, S. W., and Cheng, W. K. (2023). State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 56(1):749–780.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.
- Gallazzi, M., Biavaschi, S., Bulgheroni, A., Gatti, T., Corchs, S., and Gallo, I. (2024). A large dataset to enhance skin cancer classification with transformer-based deep neural networks. *IEEE Access*.
- Gallo, I., Rehman, A. U., Dehkordi, R. H., Landro, and Nicola (2023). Deep object detection of crop weeds: Performance of yolov7 on a real case dataset from uav images. *Remote Sensing*, 15(2):539.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- ISIC Challenge (Last accessed on 30-10-2024). Isic challenge webpage. <https://challenge.isic-archive.com>.
- Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768.
- Liu, Z., Lin, Y., Cao, Y., and Hu (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- Mao, H. H. (2020). A survey on self-supervised pre-training for sequential transfer learning in neural networks. *arXiv preprint arXiv:2007.00800*.
- O’Shea, K. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
- Paulsen, S. and Casey, M. (2023). Sequential transfer learning to decode heard and imagined timbre from fmri data. *arXiv preprint arXiv:2305.13226*.
- Redmon, J. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Rehman, A. U. and Gallo, I. (2024). Cross-pollination of knowledge for object detection in domain adaptation for industrial automation. *International Journal of Intelligent Robotics and Applications*, pages 1–19.
- Rehman, A. U., Gallo, I., and Lorenzo, P. (2023). A food package recognition framework for enhancing efficiency leveraging the object detection model. In *2023 28th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Germany, 2015, proceedings, part III 18*, pages 234–241. Springer.
- Taghizadeh, M. and Mohammadi, K. (2022). The fast and accurate approach to detection and segmentation of melanoma skin cancer using fine-tuned yolov3 and segnet based on deep transfer learning. *arXiv preprint arXiv:2210.05167*.
- Tirinzoni, A., Poiani, R., and Restelli, M. (2020). Sequential transfer in reinforcement learning with a generative model. In *International Conference on Machine Learning*, pages 9481–9492. PMLR.
- Tschandl, P. and Rosendahl (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of skin lesions. *Scientific data*, 5(1):1–9.
- Ultralytics (2024). Yolov8 release notes. <https://github.com/ultralytics/yolov8>. Available: <https://github.com/ultralytics/yolov8>.
- Wang, C.-Y., Liao, H.-Y. M., and Wu, Y.-H. (2020). Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391.
- Wang, H., Xie, S., Lin, L., and Iwamoto (2022). Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2390–2394. IEEE.
- Wang, Y., Su, J., Xu, Q., and Zhong, Y. (2023). A collaborative learning model for skin lesion segmentation and classification. *Diagnostics*, 13(5):912.