ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Leveraging Segmentation Maps to improve Skin Lesion Classification

Simone Bonechi[1], Paolo Andreini[1] and Fiamma Romagnoli[2]

1 - Department of Information Engineering and Mathematics,
University of Siena, Siena, Italy

2 - Institute of Informatics and Telematics, CNR, Pisa, Italy

Abstract.
We propose a novel approach for skin lesion classification that leverages a transformer architecture to integrate diverse clinical information (dermoscopic images, segmentation maps, and patient clinical information) for more accurate diagnosis. By incorporating binary semantic segmentation maps as input, we directly provide the model with border details critical for distinguishing between benign and malignant lesions. This integration improves classification performance compared to models that use only dermoscopic images or clinical data. To the best of our knowledge, this is the first application of segmentation maps to enhance skin lesion classification. Our experiments on the ISIC dataset yield promising results, highlighting the potential of combining advanced transformer models with multimodal data for improved dermatological diagnostics.

1 Introduction

Early diagnosis of malignant melanoma (MM) is critical but remains a challenging task in dermatology. Timely detection is essential for reducing MM mortality rates, which accounted for approximately 59,000 deaths worldwide in 2022 alone [1]. Dermoscopic examination is the primary diagnostic tool for MM, but its effectiveness is substantially influenced by the dermatologist's expertise and the subjective interpretation of dermoscopic features within melanocytic skin lesions (MSLs) [2, 3]. To assist clinicians and improve diagnostic accuracy, automated tools leveraging deep learning have emerged. While Convolutional Neural Networks (CNNs) have been widely used for skin lesion classification [4, 5], newer architectures like Vision Transformers (ViTs) offer advantages in modeling global context [6, 7, 8] and integrating multimodal data. Previous studies have focused on incorporating additional data such as macroscopic images and metadata to create multimodal inputs [9, 10, 11, 12]. In this paper, we introduce a novel approach that leverages transformer architectures to integrate dermoscopic images, binary semantic segmentation maps (lesion and background), and clinical data for enhanced skin lesion classification. By carefully designing the integration of skin lesion segmentation maps as inputs, we provide the model with crucial border details that help determine whether a lesion is benign or malignant. This approach aligns with clinicians' emphasis on lesion borders and shapes in their diagnostic process. To the best of our knowledge, we are the first to use segmentation maps as additional input to skin lesion classification models. Unlike previous studies that incorporate dermoscopic images and metadata for multimodal inputs, our method integrates segmentation maps to enhance the model's understanding of lesion morphology. In our experiments, we modified the Simple ViT [13] model to integrate dermoscopic images, segmentation maps, and patient clinical data. The results on the ISIC dataset demonstrate that this integrated approach improves classification accuracy, offering a promising direction for automated skin cancer diagnosis. The paper is organized as follows. In Section 2, the experimental setups and the Simple ViT model are described together with the ISIC dataset; then, the results are presented and discussed in Section 3. Finally, Section 4 collects conclusions and future perspectives.

2 Materials and Methods

2.1 ISIC Dataset

The International Skin Imaging Collaboration (ISIC) dataset is a comprehensive, publicly accessible repository of high-quality dermoscopic images, designed to support advancements in skin lesion analysis and melanoma detection. The dataset encompasses a wide variety of skin lesions, including benign nevi and malignant melanomas, accompanied by detailed metadata and diagnostic labels. Regularly expanded through annual ISIC challenges, the dataset now includes tens of thousands of images, ensuring consistency and quality for robust model development. In this study, we utilized images from the ISIC challenges held in 2017, 2019, and 2020. Duplicate images were removed following the methodology described in [14]. Since the 2017 and 2019 challenges were designed for multiclass classification tasks, we converted their labels to binary classification (melanoma vs. non-melanoma) to match the format of the 2020 challenge. Segmentation maps for all images were generated using the approach proposed in [15]. For validation, we randomly selected 20% of the 2020 dataset, preserving its original class distribution of approximately 10% melanomas. The remaining 2020 images were merged with those from the 2017 and 2019 datasets to form the training set. To mitigate class imbalance, melanoma cases in the training data were oversampled to equalize the number of non-melanoma images. This process resulted in a training set of 93,170 images and a validation set of 6,531 images. The test set consisted of the original ISIC 2020 test set, comprising 10,982 images without publicly available annotations. Consequently, all reported results were obtained through the official submission server for the ISIC 2020 challenge on Kaggle[1].

2.2 Simple ViT Model

Simple ViT [13] is a streamlined variant of the ViT [16] tailored for image classification. Unlike conventional CNNs, which utilize convolutional layers to model spatial hierarchies, Simple ViT employs a transformer-based architecture that

[1] https://www.kaggle.com/c/siim-isic-melanoma-classification/

represents images as sequences of patches. Each image is divided into fixed-sized patches, which are linearly embedded and treated as input tokens for the transformer. Through self-attention mechanisms, the model learns global relationships among these patches, enabling it to capture both local and long-range dependencies in the image. A distinctive feature of Simple ViT is its simplicity and modularity, with fewer design complexities compared to more sophisticated vision transformers. This model comprises a straightforward embedding layer to process patches, a series of transformer blocks for attention and feed-forward operations, and a final classification head that generates the output predictions.

2.3 Experimental Setup

This paper presents a preliminary investigation into the impact of incorporating segmentation maps as an additional input to a transformer model (Simple ViT) for classifying malignant skin lesions. Segmentation maps can enhance the model's ability to focus on the lesion area and, more importantly, provide explicit information about lesion boundaries — a critical feature for dermatological analysis. This study aims to determine whether directly supplying this boundary information improves classification performance. Although the model can independently learn lesion boundaries, explicitly providing this information can simplify the learning process and enable it to focus more effectively on capturing finer details. We evaluated two strategies for integrating segmentation maps into the transformer input (Figure 1).

1. Token level concatenation (**Token-concat**) - Image and segmentation tokens are concatenated along the token dimension, effectively doubling the total number of tokens while keeping each token's feature dimensionality fixed at 1024.
2. Feature level concatenation (**Feature-concat**) - Each image token is concatenated with its corresponding segmentation token along the feature dimension, preserving the total number of tokens but increasing each token's feature dimensionality to 2048.

In both approaches, segmentation patches are generated identically to image patches, with positional encoding applied to ensure that corresponding image and segmentation patches share the same position identifiers. To assess the impact of incorporating segmentation maps, we compared the two proposed approaches against two baseline Simple ViT models trained exclusively on image data. We conducted two experiments to compare networks with the same number of parameters. In the first, we compared **Baseline1024** with a ViT model using the Token-concat strategy, keeping the input embedding dimension at 1024. In the second, using the Feature-concat strategy, we concatenated 1024-dimensional embeddings of images and segmentation maps to create a 2048-dimensional input. This network was compared with **Baseline2048**, which also has a 2048 input dimension, ensuring both models had the same number of parameters.

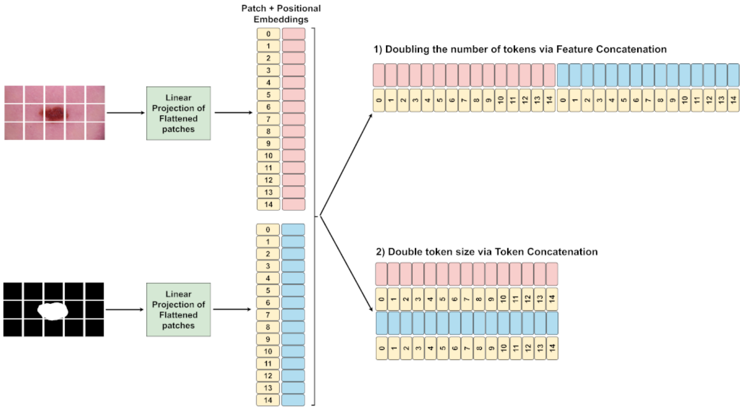
Patch + Positional Embeddings
Linear Projection of Flattened patches
Linear Projection of Flattened patches
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1) Doubling the number of tokens via Feature Concatenation
2) Double token size via Token Concatenation

Fig. 1: Concatenation strategies employed to combine images and segmentation maps.

Finally, we aimed to assess the impact of incorporating patient metadata into the model (Figure 2). To integrate metadata, we adapted the Simple ViT architecture by adding an initial three-layer Multilayer Perceptron (MLP) dedicated to processing the metadata. Starting from a size-twelve input the MLP gradually increases the feature dimensionality to match the transformer's output dimension, equal to its input dimension. This representation is then concatenated with the transformer's encoding. The resulting combined representation is then passed through the final classification layers. Using this architecture, we evaluated the impact of metadata in all the previously described configurations: Simple ViT model integrated with segmentation in both Token-concat and Feature-concat setups (**Token-concat + Metadata** and **Feature-concat + Metadata**), as well as in the two baseline models without segmentation (**Baseline1024 + Metadata** and **Baseline2048+Metadata**). For all experiments, we adopted the same hyperparameters used in the original Simple ViT study [13][2]. Furthermore, various augmentation techniques, including rotation, flipping, and adjustments to brightness, contrast, and color, were applied during

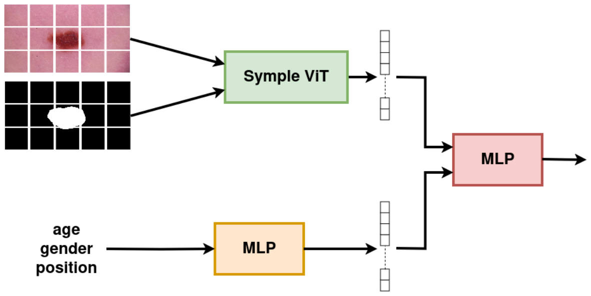

Fig. 2: Incorporating patient metadata in the model.

[2] Image size: 256 × 256; Patch size: 16; Transformer depth: 6; Transformer heads: 8; Final

training to expand the dataset. The validation set was utilized for early stopping to prevent overfitting.

3 Results

In Table 1 we present the results obtained using the experimental setup detailed in Section 2.3. Since the test set labels are not publicly available, the results for this set were obtained through the ISIC 2020 challenge submission server.

	Test set		Validation set
Setup	**AUC private**	**AUC public**	**AUC**
Baseline1024	0.8505	0.8706	0.8644
Baseline1024 + Metadata	0.8441	0.8727	0.8654
Token-concat	**0.8663**	0.8719	0.8701
Token-concat + Metadata	0.8629	**0.8772**	**0.8730**

(a) Performance with token dimension equal to 1024

	Test set		Validation set
Setup	**AUC private**	**AUC public**	**AUC**
Baseline2048	0.8546	0.8592	0.8728
Baseline2048 + Metadata	0.8615	0.8680	0.8728
Feature-concat	0.8634	0.8843	0.8774
Feature-concat + Metadata	**0.8694**	**0.8888**	**0.8905**

(b) Performance with token dimension equal to 2048

Table 1: Comparing baseline models with those incorporating segmentation and metadata using Area Under the Receiver Operating Characteristic Curve (AUC).

The results demonstrate that incorporating segmentation maps improves performance compared to baseline models in both configurations (Token-concat and Feature-concat). Specifically, the concatenation of segmentation maps along the feature dimension (Token-concat configuration) is the most effective. This may be because this approach directly correlates image features with segmentation patches at corresponding locations, while token-based concatenation in the TLC setup requires the model to infer these positional correspondences through positional embeddings.

The same trend is evident when metadata is included during training, further validating the benefit of using segmentation maps.

4 Conclusions

In this study, we evaluate the impact of incorporating segmentation maps as additional input to a Vision Transformer (ViT) for skin lesion classification. We adapted the Simple ViT model to integrate information from dermoscopic images, segmentation maps, and patient clinical data. Experiments on the ISIC dataset show that segmentation maps improve classification performance compared to models using only dermoscopic images or a combination of images and

MLP dimension: 2048; Learning rate: 0.001; Batch size: reduced to 110 to accommodate the memory constraints of the employed NVIDIA GeForce RTX 4090 GPU.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

clinical data. By providing explicit border details, this approach aligns with clinicians' emphasis on such features and highlights the potential of leveraging multimodal data to enhance diagnostic accuracy in dermatology. To our knowledge, this is the first study to utilize segmentation maps as input for skin lesion classification. Future work could explore incorporating metadata directly into the transformer's input for a unified processing framework and examine the role of segmentation maps under different models' hyperparameters, such as image and patch size.

References

[1] Freddie Bray et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

[2] Hans Skvara et al. Limitations of Dermoscopy in the Recognition of Melanoma. *Archives of Dermatology*, 141(2):155–160, 02 2005.

[3] Ana Maria Forsea et al. The impact of dermoscopy on melanoma detection in the practice of dermatologists in europe: results of a pan-european survey. *Journal of the European Academy of Dermatology and Venereology*, 31(7):1148–1156, 2017.

[4] Andre Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[5] S Lee et al. Augmented decision-making for acral lentiginous melanoma detection using deep convolutional neural networks. *Journal of the European Academy of Dermatology and Venereology*, 34(8):1842–1850, 2020.

[6] Xinzi He et al. Fully transformer network for skin lesion analysis. *Medical Image Analysis*, 77:102357, 2022.

[7] Chao Xin et al. An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, 149:105939, 2022.

[8] Rahmat Izwan Heroza et al. Enhancing skin lesion classification: A self-attention fusion approach with vision transformer. In *Annual Conference on Medical Image Understanding and Analysis*, pages 309–322. Springer, 2024.

[9] Jordan Yap et al. Multimodal skin lesion classification using deep learning. *Experimental dermatology*, 27(11):1261–1267, 2018.

[10] Jeremy Kawahara et al. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[11] Simone Bonechi et al. Fusion of visual and anamnestic data for the classification of skin lesions with deep learning. In *New Trends in Image Analysis and Processing–ICIAP*, pages 211–219. Springer, 2019.

[12] Linda Tognetti et al. A new deep learning approach integrated with clinical data for the dermoscopic differentiation of early melanomas from atypical nevi. *Journal of Dermatological Science*, 101(2):115–122, 2021.

[13] Lucas Beyer et al. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

[14] Bill Cassidy et al. Analysis of the isic image datasets: Usage, benchmarks and recommendations. 75:102305, 2022.

[15] Simone Bonechi. Isic_wsm: Generating weak segmentation maps for the isic archive. *Neurocomputing*, 523:69–80, 2023.

[16] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. Cited by: 7975.