Attention-Based Color Fusion in Dermoscopy AI

Alma Pedro[1,2], Ana M. Cabanas[2,3], Atsuko Galaz[1], Soledad Vidaurre[4], Cristian Navarrete-Dechent[2,5], Domingo Mery[1,2]

[1] Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
[2] Millennium Institute for Intelligent Healthcare Engineering, Santiago, Chile
[3] Physics Department, Universidad de Tarapacá, Arica, Chile
[4] Cienbio SpA, Santiago, Chile
[5] Department of Dermatology, School of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

Introduction

Skin cancer is among the most common cancers worldwide, and melanoma, though a minority of cases, drives most skin-cancer mortality, so early and accurate diagnosis is decisive. Dermoscopy and deep learning have improved automated classification, yet most CNN and transformer models learn from raw pixels and ignore the structured colour reasoning clinicians apply, comparing the lesion with the surrounding skin. We investigate whether making this colour explicit, as compact CIELAB metadata, and fusing it with the image improves seven-class lesion classification while remaining clinically interpretable. Figure 1 shows the framework evaluated in this study.

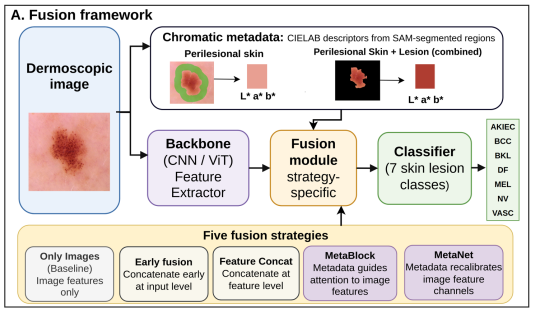

Figure 1. Fusion framework. A CNN/ViT backbone encodes the dermoscopic image; CIELAB descriptors from SAM-segmented regions (perilesional skin, or lesion + perilesional combined) are combined through a strategy-specific fusion module before the 7-class classifier. The five fusion strategies appear at the bottom.

Data & Methods

Colour from the lesion and the perilesional skin is summarised in CIELAB (L* a* b*), a perceptually uniform space where equal numerical distances correspond to similar perceived colour differences. Descriptors (L*= lightness (black-white), a*= green-red, b*= blue-yellow) were extracted with SAM-based (Segment Anything Model) segmentation of two regions, the lesion and the perilesional skin, on HAM10000 (9,958 images; 7 classes: AKIEC, BCC, BKL, DF, MEL, NV, VASC). Three backbones (ViT-B/32, ResNet50, VGG11) were trained under five fusion strategies and evaluated with five-fold stratified cross-validation (F1, BACC, AUC).

Five fusion strategies, how colour is combined with the image:

Only-Images	Baseline — image features only; metadata ignored.
Early fusion	Colour appended to the input, before the backbone.
Feature Concat	Colour joined to image features, before the classifier.
MetaBlock	Attention lets colour reweight the image feature maps.
MetaNet	Colour recalibrates entire feature channels (channel-wise scaling).

Results

We evaluated three backbones (ViT-B/32, ResNet50 and VGG11) under five fusion strategies and two colour-metadata settings, perilesional skin alone, or lesion + perilesional skin combined, using five-fold stratified cross-validation on HAM10000. Performance is reported as macro-F1, balanced accuracy (BACC) and area under the ROC curve (AUC). Because the transformer backbone was consistently the strongest, Table 1 summarises the ViT-B/32 results across every fusion × metadata combination, with the remaining backbones following the same trends at lower absolute scores.

Fusion	Metadata	F1	BACC	AUC
Only-Images	—	0.829	0.719	0.954
Early fusion	Perilesional	0.833	0.719	0.953
Feature Concat	Perilesional	0.836	**0.737**	0.958
MetaBlock	Perilesional	0.816	0.678	0.951
MetaNet	Perilesional	0.831	0.724	0.954
Feature Concat	Lesion + Skin	0.837	0.733	0.961
MetaBlock	**Lesion + Skin**	**0.844**	0.736	**0.962**
MetaNet	Lesion + Skin	0.836	0.725	0.960

The best overall performance was achieved by ViT-B/32 combined with MetaBlock and lesion + perilesional chromatic metadata (F1 = 0.844, BACC = 0.736, AUC = 0.962), compared with an image-only baseline of BACC = 0.719 and AUC = 0.954 (Figure 1B). For MetaBlock specifically, switching from perilesional-only to lesion + perilesional metadata produced the largest gain observed (+5.8 BACC points and +1.2 AUC). Early fusion barely moved the needle; gains accrued mainly to the attention-based methods, and chiefly to the transformer backbone, while CNN backbones improved less.

Limitations & Future Work

Our findings derive from a single dataset (HAM10000) and need external, multi-centre validation. Class imbalance may still affect minority-class metrics despite stratified cross-validation, and because colour depends on illumination, optics and skin tone, future work should test robustness across devices and diverse phototypes, and explore descriptors beyond CIELAB.

Conclusion

Lesion and perilesional skin carry complementary, non-redundant colour information, but models exploit it only when attention weighs colour against image content — simple concatenation is not enough. The best setup pairs a transformer backbone with metadata-guided attention (MetaBlock) and colour from both regions. Because CIELAB is perceptually meaningful, the approach stays clinically interpretable.

References

[1] Tschandl, Philipp; Rosendahl, Cliff; Kittler, Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 2018, 5:180161.
[2] Kirillov, Alexander; Mintun, Eric; Ravi, Nikhila; Mao, Hanzi; Rolland, Chloé; Gustafson, Laura; Xiao, Tete; Whitehead, Spencer; Berg, Alexander C.; Lo, Wan-Yen; Dollár, Piotr; Girshick, Ross. Segment Anything. ICCV, 2023, 4015–4026.
[3] Pacheco, Andre G. C.; Krohling, Renato A. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. IEEE J. Biomedical and Health Informatics, 2021, 25(9):3554–3563.
[4] Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil. An image is worth 16×16 words: Transformers for image recognition at scale. ICLR, 2021.

Acknowledgements. A.P., D.M., C.N.-D., and A.M.C. acknowledge support from the Millennium Institute for Intelligent Healthcare Engineering ICN2021_004. A.P. and A.M.C. acknowledge support from the National Center for Artificial Intelligence (CENIA) FB210017 Basal ANID. A.M.C. acknowledges support from Fondecyt Regular 1251081