Fusion of U-Net++ and ResNet50 Models for Melanoma Diagnosis from Dermoscopic Images

Kilian Zell, 287515

Abstract

This report relates the development and successful implementation of a fusion model that combines a custom encoder-decoder neural network with a pre-trained residual network for skin lesion segmentation and classification from dermoscopic images. Designed specifically for melanoma diagnosis, the resulting model reached an overall accuracy of 94%.

1. Introduction

Skin cancer is the most prevalent type of cancer, and while melanoma constitutes only around 1% of them, it is responsible for the majority of skin cancer-related death.(1) Currently, melanoma diagnoses rely primarily on the visual examination of skin lesions by dermatologists.(2) Unfortunately, this kind of diagnosis is prone to subjectivity and errors.(3) This significantly impact patients' well-being, as an early-stage misdiagnosis can substantially diminish overall chances of survival.

In this context, deep learning has the potential to significantly enhance the consistency and accuracy of diagnosis while remaining entirely non-invasive. In this work, a model specifically designed to classify skin lesions as 'melanoma' or 'non-melanoma' is proposed. In unsegmented dermoscopic images, the borders of a lesion and skin texture can lead to unwanted feature extraction, affecting the overall performance of classification. To overcome this challenge, the proposed model first extract the Region of Interest (ROI) form the input image using a U-Net++ inspired model so that only relevant features are considered for diagnosis. The resulting segmented region is then fed into a pre-trained ResNet50 calibrated for binary classification to obtain the final prediction.(see Fig. 2)

The entire workflow is trained and evaluated using the HAM10000(4) dataset, consisting of 10'015 dermoscopic images along with their corresponding binary mask images and entries indicating gold standard malignant status. (see Fig. 1)

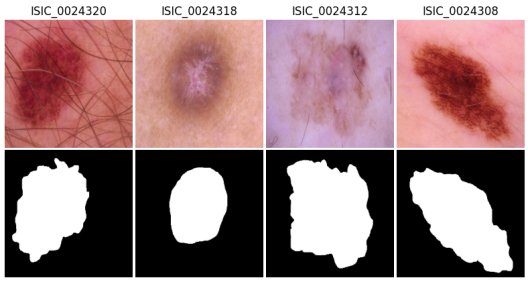

Figure 1. Random examples of skin lesion images (top) and their corresponding masks (bottom) from the HAM10000 dataset after resizing to 256 * 256.

2. Proposed Combinatory Model

2.1. Input dataset

The HAM10000(4) dataset was used for to both train and test the proposed model. It comprises 10'015 dermoscopic images of skin lesions, all in RGB format and with a .JPG extension. The corresponding ground truth masks are in grayscale and .PNG format. Both the original images and their ground truth counterparts have dimensions of 600 x 450 pixels. The dataset encompasses a total of seven skin disease labels, including melanoma (MEL), vascular lesions (VASC), benign keratosis (BKL), and others. Since this work focuses on binary classification for melanoma diagnosis, only the melanoma label was considered.

2.2. Unet++ Inspired Architecture

The semantic segmentation part of the proposed model is inspired by a U-Net++ architecture. This encoder-decoder model is a variant of the original U-Net architecture and is designed to capture more contextual information by incorporating nested skip connections in the decoding part. Compared to the original model, which typically uses only direct connections between corresponding encoder and decoder layers, the U-Net++ architecture incorporates skip connections at different levels of the encoding hierarchy. Specifically, each decoding block of the model is connected not only to the corresponding encoding block but also to the previous one (see Fig. 3).

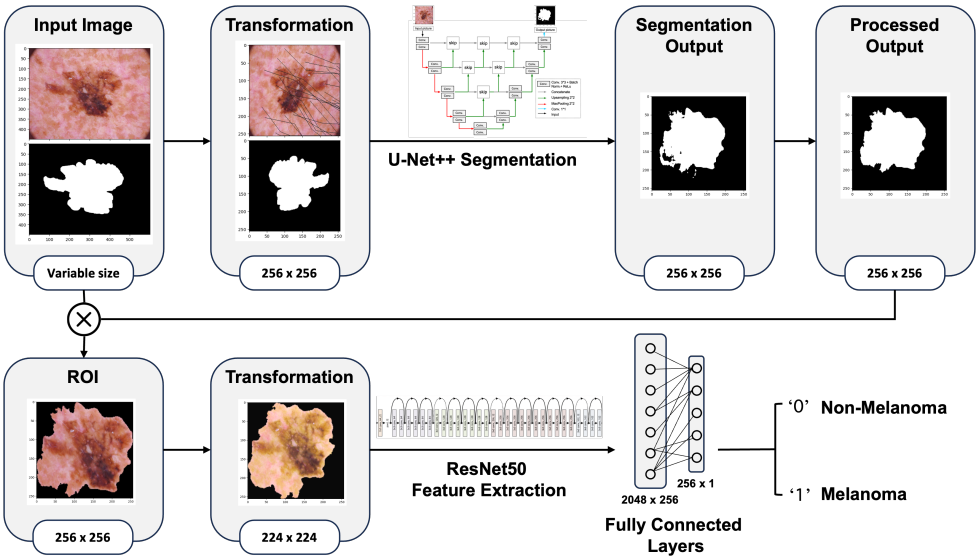

Figure 2. Workflow of the proposed combination model.

In the first encoding layer of the model, an input image of size 256 * 256 is passed through a convolutional block consisting of two consecutive convolutional layers with a kernel size of 3 * 3, padding of 1, and ReLU activation. To mitigate issues like internal covariate shift, each convolutional layer is followed by batch normalization. Max-pooling is then applied with a kernel size of 2 * 2 to downsample the spatial dimensions of the feature maps by a factor of two. The resulting object is then passed to three additional encoding layers of similar structure to achieve a final minimal resolution of 16 * 16 at the bottleneck. At this point, the decoder part of the model increases the size of the feature map using 2 * 2 transposed convolutions in between four convolution blocks. At each encoding step, features extracted from the nested skip connections are concatenated to the feature map before convolution. In the last step, the resulting feature map of spatial dimension 256 * 256 is convoluted with a kernel size of 1 * 1 to give a probability map of the same spatial size as the input image. Finally, a sigmoid activation function is applied to obtain the pixel-wise class prediction.

2.3. Pre-Trained ResNet50

Following segmentation, the proposed workflow employs a binary classification model built upon the ResNet50 architecture sourced from the torchvision models library(5). The feature extraction layers comprise this ResNet50 architecture, which was pre-trained on the ImageNet(6) dataset and excluded from its terminal prediction head. By leveraging the knowledge encapsulated during pre-training on a diverse set of images, these layers should rapidly capture hierarchical features from the input images. For prediction, a custom fully connected layer, consisting of two linear layers

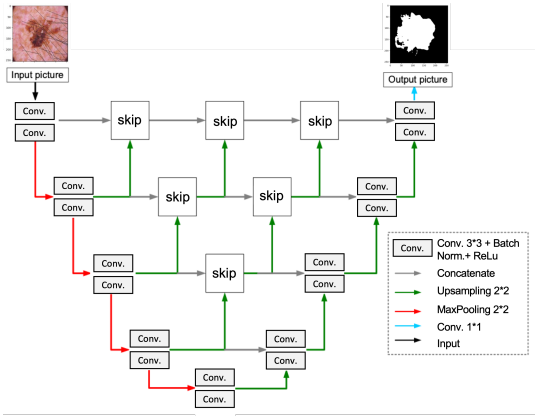

Figure 3. Proposed architecture for the semantic segmentation model inspired by Unet++.

with ReLU activation, dropout, and batch normalization, is followed by sigmoid activation. Finally, the model outputs a binary prediction, where the positive class corresponds to 'melanoma.'

3. Experimental Settings

3.1. Data Processing

3.1.1. Segmentation

The input data is initially split into training (80%) and testing (20%) sets. Before applying segmentation, both the original and ground truth images are resized to 256 * 256. To prevent over-fitting, stochastic transformations are applied to the training images. In addition to classical augmenta-

tion techniques such as flips, rotations or zooms, mosaic transformation—consisting of combining four images by placing each half-size resized image in one corner—and cutmix—consisting of randomly integrating a patch of another image into the original image—are applied. (see Fig. 4)

Furthermore, in Fig. 1, the first image (ISIC0024320) displays a lesion that is covered by hairs. These hairs may adversely affect the model's performance and pose a significant challenge for segmentation models. To address this issue, artificial hairs are stochastically introduced using a custom methode, as depicted in Fig. 4 (ISIC0024308).

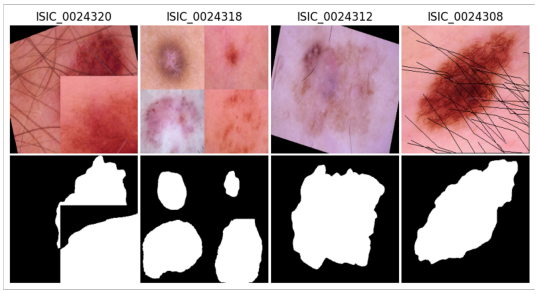

Figure 4. Examples of random transformations applied to the training dataset before segmentation. (ISIC0024320: random rotation and cutmix, ISIC0024318: mosaic, ISIC0024312: horizontal flip and random rotation, ISIC0024308: vertical flip and artificial hairs)

3.1.2. ROI Extraction

After segmentation, the generated mask predictions undergo refinement through an image processing technique. This technique involves detecting contours, and subsequently filling the largest detected area with positive predictions, while the remaining areas are set to zero. This process ensures that only the primary skin lesion is considered for the classification model and effectively reduces false negative instances within the positively detected area.

The Region of Interest (ROI) is subsequently extracted by multiplying the original image with the processed segmentation prediction. The resulting product retains the original image while setting the background (skin around the lesion) to zero. Finally, the extractor identifies the bounding box coordinates encapsulating the ROI and crops the smallest squared window possible around the lesion. (see Fig. 5)

3.1.3. Classification

Similar to the segmentation step, the input data is initially split into training (80%) and testing (20%) sets. To ensure compatibility with the pre-trained ResNet50, the ROI images are resized to 224 * 224. Various transformations, such as horizontal and vertical flips, random rotations, zooms, color jitters—consisting of random modifications of brightness, contrast, saturation, and hue—and random cutouts, are applied to the training images to prevent over-fitting.

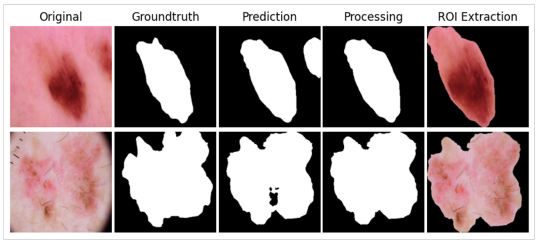

Figure 5. Examples of region of interest (ROI) extraction from segmentation predictions.

The HAM10000 dataset demonstrates significant class imbalance, with only about 11% of samples labeled as 'melanoma.' To address this imbalance and improve the model's generalization, an up-sampling strategy was implemented. Training samples are assigned weights inversely proportional to their class sample counts. During data loading, each training sample has a probability of being drawn in a batch proportional to its assigned weight. Additionally, samples are allowed to be drawn again in subsequent batches. This approach ensures a homogeneous representation of both classes throughout the training process.

3.2. Procedure

The U-Net++ inspired model is trained using binary cross-entropy dice loss, while the ResNet50 is trained using the binary cross-entropy loss provided by the PyTorch library. Various hyperparameters, including dropout and weight decay, were fine-tuned for both models using a random search approach. Both models were trained for 40 epochs on an Apple M3 Pro GPU with batch sizes of 32 for both segmentation and classification. The training loss curves for both models are depicted in Fig. 6.

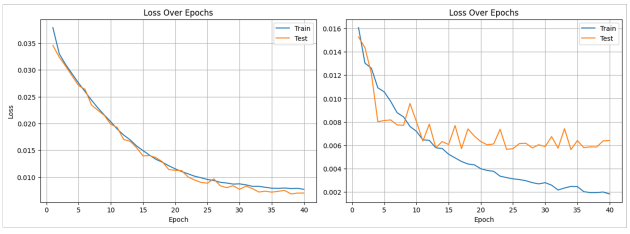

Figure 6. Training and testing loss for the segmentation (left) and the classification (right) model after 40 epochs.

4. Results and Discussion

In this work, pixel-wise accuracy and dice score were employed as performance metrics for segmentation. For the classification model, the evaluation encompassed accuracy, specificity, sensitivity, precision, and F1-score.

For segmentation, two models were considered: a simple CNN with three convolutional layers and two fully connected layers, a classical U-Net architecture, and the aforementioned U-Net++ inspired model. Analysis of Table 1 reveals that, in terms of accuracy and dice coefficient, the most promising model is the U-Net++ inspiration, achieving an accuracy of 95.56% and a dice score of 91.95%. Furthermore, the introduction of the proposed post-processing method slightly enhanced the overall performance of the segmentation model, resulting in a final performance of 95.66% accuracy and 92.10% dice score.

Table 1. Segmentation Results

MODEL	ACC.	DICE
RANDOM	49.78%	31.09%
CONST. NEG.	73.28%	0.01%
SIMPLE CNN	94.77%	89.51%
UNET	95.21%	90.90%
UNET++	95.56%	91.95%
UNET++ & PROC.	**95.66%**	**92.10%**

For classification, several pre-trained ResNet architectures were considered. To mitigate any bias from the segmentation, classification models were initially tested on ROIs extracted from the ground-truth masks rather than the segmentation model output (labelled 'B.' in Table 2). As shown in Table 2, ResNet50 emerged as a promising model, achieving an accuracy of 94.21%, specificity of 96.63%, sensitivity of 75.11%, and F1-score of 74.45%. When evaluated on ROI images predicted by the segmentation model, the selected classification model demonstrated a final score of 94.06% accuracy, 96.51% specificity, 74.67% sensitivity and 73.85% F1-score, indicating that segmentation errors had negligible impact on the final predictions.

Table 2. Classification Results

MODEL	ACC.	SPEC.	SENS.	F1
RANDOM	50.12%	49.04%	53.33%	19.18%
CONST. NEG.	88.77%	100%	0.00%	0.00%
RESNET34 B.	91.36%	93.48%	74.67%	66.01%
RESNET50 B.	94.21%	96.63%	75.11%	74.45%
RESNET101 B.	94.66%	98.09%	67.56%	73.97%
RESNET50	**94.06%**	**96.51%**	**74.67%**	**73.85%**

Fig. 7 displays the output of the proposed fusion model at different stages for selected test samples, along with their corresponding ground-truth.

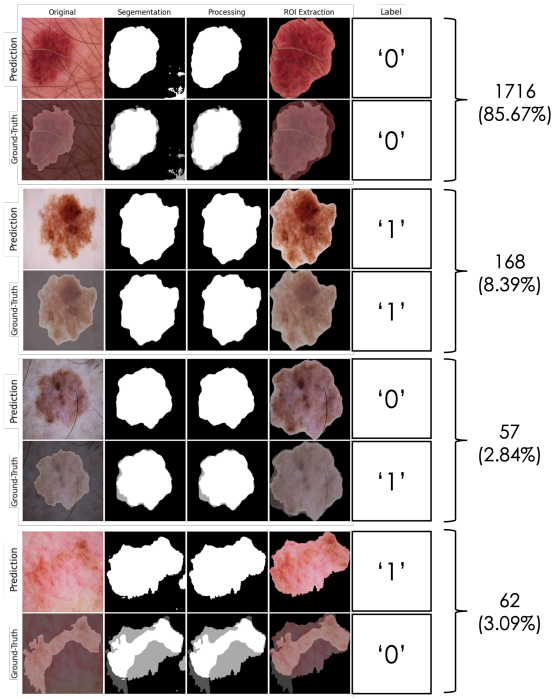

Figure 7. Output of the proposed fusion model for certain test set images is presented at different stages alongside corresponding ground-truth labels. The labels '0' and '1' denote non-melanoma and melanoma, respectively. The numbers on the right indicate the number of true negatives, true positives, false positives, and false negatives predicted for the test set from top to bottom.

5. Conclusion

This work presents a strategy that combines a segmentation and classification model for melanoma diagnosis from dermoscopic images. For this purpose, an encoder-decoder model inspired by the U-Net++ architecture was developed and trained on skin lesion. The output of this model was then utilized to isolate Region of Interest (ROI) images, which were subsequently used as inputs in a classification model based on a pre-trained ResNet50 architecture calibrated for binary predictions. The entire workflow demonstrated a final accuracy of 94.06% and a F1 score of 73.85%.

In addition to its technical contributions, the proposed strategy has noteworthy clinical implications, potentially serving as a valuable second opinion tool for dermatologists in melanoma diagnosis. Furthermore, future enhancements could involve integrating patient metadata, such as age or gender, into the classification head, thereby providing a more comprehensive and personalized diagnostic approach. This integration could further refine the model's insights and contribute to its utility in clinical settings.

Table 3. Literature Review

REF.	METHODE	DATASET	ACC.	MEL SPEC.	MEL SENS.	MEL F1
MAJTNER ET AL. (2016)	CNN & RSURF FEATURES	ISIC2016	82.60%	89.81%	53.33%	-
ZHANG ET AL. (2019)	(ARL-CNN) MODEL	ISIC2017	87.40%	89.60%	65.80%	-
MASNI ET AL. (2020)	RESNET50	HAM10000	87.74%	89.34%	61.90%	65.39%
ALSAHAFI ET AL. (2023)	SKIN-NET	ISIC2019	94.65%	99.20%	50.89%	56.20%
VATSALA ET AL. (2023)	U-NET & CNN	HAM10000	97.96%	97.00%	62.22%	53.33%
PROPOSED MODEL	**U-NET++ & RESNET50**	**HAM10000**	**94.06%**	**96.51%**	**74.67%**	**73.85%**

References

[1] American Cancer Society. (n.d.). Key statistics for melanoma skin cancer. Retrieved from https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html

[2] Davis, E. L., Shalin, S. C., Tackett, A. J. (2019). Current state of melanoma diagnosis and treatment. Cancer Biology and Therapy, 20(11), 1366–1379. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6804807

[3] Grant-Kels, J. M., Bason, E. T., Grin, C. M. (1999). The misdiagnosis of malignant melanoma. Journal of the American Academy of Dermatology, 40(4), 539–548. https://doi.org/10.1016/s0190-9622(99)70435-4

[4] Tschandl, P., Rosendahl, C., Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161. doi:10.1038/sdata.2018.161

[5] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385 [cs.CV]. https://doi.org/10.48550/arXiv.1512.03385

[6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet Large Scale Visual Recognition Challenge. Retrieved from http://www.image-net.org/challenges/LSVRC/2010/

6. Annex A: Literature Review

In the past, numerous studies have utilized deep learning algorithms to analyze skin lesions from dermoscopic images (see Table 3). While some authors focused on segmentation-based approaches and others on classification-based approaches, only a few integrated both approaches into their pipeline.

Furthermore, the vast majority of classification-oriented publications considered multi-label classification (7 to 9 lesion categories). While this might be useful in the context of education or research, the result of those studies demonstrate moderate sensitivity regarding the melanoma class (MEL) potentially resulting in misleading clinical diagnosis.

The proposed work focuses exclusively on melanoma diagnosis and adopts a U-Net++ and RestNet50 fusion model for segmenting and classifying skin lesions.

7. Annex B: Metadata Integration

In anticipation of future potential clinical applications, the integration of patient metadata, such as age and sex, into the proposed model was investigated. The ISIC2019 dataset, which compiles 25'331 dermatoscopic images in .JPG format along with corresponding patient metadata entries was considered a suitable candidate to train and test such model. Please note that this aspect of the project was successfully implemented but has not undergone complete testing.

To encode the metadata, a straightforward mapping method is employed. Specifically, each unique element in a given metadata feature is assigned a distinct value starting from '1'. Empty elements are default-mapped to '0'. The assigned values are then normalized to achieve final feature values ranging between 0 and 1. The integration of these mapped metadata features into the classification model is facilitated through the introduction of an additional set of fully connected layers dedicated to processing the metadata features. Each layer comprises a linear transformation followed by ReLU activation and batch normalization. The output of these layers is then concatenated with the features extracted from the ResNet architecture. Ultimately, a last fully connected prediction head is applied, followed by sigmoid activation to obtain the final predictions.

Table 4. Preliminary Results of Metadata Integration. ResNet50 incorporating metadata was trained for 10 epochs on ISIC2019.

MODEL	ACC.	SPEC.	SENS.	F1
RANDOM	50.95%	51.63%	51.79%	26.79%
CONST. NEG.	82.93%	100%	0.00%	0.00%
RESNET50	94.06%	96.51%	74.67%	73.85%
RESNET50 META	86.50%	89.29%	72.95%	64.85%