Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

NeviUNet: An Efficient Attention-Enhanced Network for Skin Lesion Segmentation in Total Body Images

Karolina Dzwonkowska
Lodz University of Technology
Piotr Marcin Szczypiński
piotr.szczypinski@p.lodz.pl

Lodz University of Technology

Article

Keywords:

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10556854/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

NeviUNet: An Efficient Attention-Enhanced Network for Skin Lesion Segmentation in Total Body Images

Karolina Dzwonkowska[1] **and Piotr M. Szczypiński**[1,*]

[1]Lodz University of Technology, Stefana Żeromskiego 116, 90-924 Lodz, Poland
[*]piotr.szczypinski@p.lodz.pl

ABSTRACT

This paper addresses the challenge of efficient and accurate automated segmentation of skin lesions in total body photography acquired under clinical conditions. The study utilizes an image dataset captured via a prototype Dermo Studio total-body scanner designed for multi-perspective documentation of the patient's skin surface. To balance segmentation accuracy with computational efficiency, a modified deep neural network architecture based on the U-Net model is proposed. The network integrates selected design enhancements inspired by ResUNet++, including the Mish activation function, Attention Gates, and SE blocks, while omitting computationally expensive residual connections and ASPP modules. To mitigate severe class imbalance, a composite loss function combining Focal Tversky Loss and BCE was employed. Quantitative evaluation on an independent test dataset demonstrated the superiority of the proposed approach, securing a mean DSC of 0.729 and significantly outperforming the baseline configurations. The proposed architecture features low inference latency and optimized hardware resource allocation, confirming its suitability for deployment in real-time clinical medical devices.

Introduction

The skin, the largest organ of the human body, performs protective, regulatory, and sensory functions while also reflecting overall health status. Its structure reveals metabolic, immunological, and environmental processes that contribute to the formation of pigmented lesions, most commonly melanocytic nevi arising from the proliferation of pigment cells in the epidermis or dermis[1]. Their appearance and distribution depend on genetic predisposition and external factors, particularly ultraviolet radiation, which may induce photodamage and DNA mutations[2,3]. Skin lesions exhibit substantial heterogeneity, ranging from common benign nevi typically characterized by symmetry and uniform pigmentation to atypical (dysplastic) nevi presenting architectural disorder, cytological atypia, and irregular borders[4]. The presence of atypical lesions is associated with an increased risk of melanoma, which accounts for more than 57,000 deaths annually worldwide among approximately 325,000 newly diagnosed cases[5]. Early detection significantly improves survival rates, approaching 98%[6]. Consequently, long-term monitoring plays a crucial role in clinical practice, particularly for lesions that are difficult to assess during a single examination, such as amelanotic or hypomelanotic melanomas, where diagnosis often depends on identifying vascular or structural changes over time[7].

Recent advances in digital dermoscopy and total-body photography have created new opportunities for systematic monitoring of pigmented skin lesions. Total body imaging systems enable acquisition of standardized photographs of the entire skin surface from multiple perspectives, facilitating longitudinal comparisons between examinations and supporting the detection of subtle lesion changes that may indicate malignant transformation[1,7]. Such systems are particularly valuable in high-risk patients and in cases where lesion evolution provides diagnostically relevant information. Additional challenges arise in lesions with atypical morphology, including spitzoid lesions, which may mimic melanoma due to rapid growth, heterogeneous pigmentation, and mixed structural features[2,4].

In parallel with advances in image acquisition technologies, artificial intelligence has become an increasingly important tool for the automated analysis of biomedical images, including skin lesion imaging. Regardless of the specific analytical approach, segmentation remains a fundamental stage of the processing pipeline. Accurate delineation of lesion boundaries enables reliable extraction of morphological and color features, facilitates longitudinal assessment of lesion evolution, and supports subsequent diagnostic decision-making[8–10].

The introduction of deep learning, particularly convolutional neural networks (CNNs), significantly improved robustness and performance in skin lesion segmentation. The U-Net architecture, based on an encoder–decoder structure with skip connections, has become a standard in biomedical image segmentation due to its ability to capture both global context and fine image details[11]. Subsequent developments introduced residual connections[12], multi-scale processing[13], and dilated convolutions[14], attention mechanisms[15], and specialized loss functions designed to address class imbalance and improve boundary delineation[16]. These advances have substantially improved segmentation accuracy, particularly in challenging scenarios involving low contrast lesions, irregular boundaries, and severe class imbalance.

Despite these advances, a key challenge remains the trade-off between segmentation accuracy and computational efficiency. Highly complex architectures, such as ResUNet++[17], often achieve strong predictive performance at the cost of increased computational requirements, limiting their applicability in resource-constrained or real-time clinical environments.

Therefore, this study addresses the problem of efficient and accurate segmentation of skin lesions in total body images. We propose NeviUnet, a modified U-Net-based architecture that incorporates selected design elements inspired by ResUNet++[17] while maintaining a substantially lower computational complexity. The model is evaluated on a dataset of total body images acquired from five patients, covering multiple anatomical regions and imaging perspectives. The objective is to assess whether the proposed approach enables reliable and reproducible segmentation of skin lesions under challenging real-world imaging conditions while preserving computational efficiency suitable for clinical deployment.

Related work

Accurate segmentation of skin lesions from dermoscopic and clinical images is a fundamental step in computer-aided diagnosis systems for melanoma and other dermatological conditions. However, it remains a challenging task due to significant variability in lesion appearance, low contrast between lesion and surrounding tissue, irregular boundaries, and the presence of artefacts such as hair, illumination variations, and imaging noise[18]. Early approaches relied heavily on classical image processing techniques, often combined with dedicated preprocessing steps to reduce artefacts prior to segmentation.

A widely adopted preprocessing method is the DullRazor algorithm, originally introduced by Lee et al., which removes hair structures using morphological filtering followed by interpolation[19]. Although DullRazor significantly improves segmentation quality in hair-occluded images, it is primarily effective for dark, linear hair and may fail in the presence of light-coloured or irregular artefacts. Subsequent methods, such as E-shaver[20] and SharpRazor[21], extended this approach by incorporating orientation analysis and multi-scale filtering, achieving improved robustness. More recent research has explored learning-based artefact removal, including variational autoencoder (VAE)-based techniques that reconstruct hair-free images directly from data[22]. Nevertheless, these methods introduce additional computational complexity and may affect texture fidelity, indicating that reliable artefact handling remains an open challenge.

The introduction of deep learning, and in particular the U-Net architecture[11], marked a major shift in skin lesion segmentation. U-Net and its derivatives employ an encoder–decoder structure with skip connections, enabling effective learning of hierarchical and spatially precise features. Numerous studies have demonstrated the effectiveness of U-Net-based models on benchmark datasets such as ISIC (International Skin Imaging Collaboration https://www.isic-archive.com). However, as highlighted in recent surveys, convolutional neural networks (CNNs) inherently focus on local features, limiting their ability to capture global contextual information, which is essential for accurate segmentation of lesions with diffuse or irregular boundaries[18,23].

In response to these limitations, recent research has focused on enhancing U-Net architectures with additional components such as residual connections, dense feature propagation, and attention mechanisms. Attention-based models improve feature representation by selectively emphasizing relevant spatial and channel information, leading to better boundary delineation and segmentation accuracy. For instance, attention-enhanced U-Net variants and hybrid CNN architectures have demonstrated improved performance across multiple datasets, particularly in the presence of class imbalance and low contrast[24,25]. Nevertheless, these improvements often come at the cost of increased model complexity, which may lead to overfitting and reduced generalization when training data are limited.

Over the last five years, a notable trend has been the integration of transformer-based architectures with CNN models to address the limitations of purely convolutional approaches. Transformers introduce self-attention mechanisms that enable modeling of long-range dependencies and global context. Hybrid architectures such as CTH Net[26] and TESL Net[27] combine CNN encoders with transformer modules, allowing simultaneous extraction of local features and global relationships. These models have achieved state-of-the-art performance on standard dermoscopic datasets, demonstrating improved segmentation accuracy and robustness. However, they are typically computationally demanding and require large annotated datasets, which may limit their applicability in real-time or resource-constrained clinical environments.

Despite these advances, several fundamental challenges remain. Data scarcity and class imbalance continue to hinder model training and generalization, while variability in imaging conditions—such as different acquisition devices, illumination settings, and patient populations—reduces cross-dataset robustness[18,23]. Moreover, improvements in quantitative metrics such as Dice-Sørensen coefficient (or Intersection over Union – IoU) do not always translate into clinically meaningful gains, particularly in terms of boundary accuracy and reproducibility. Consequently, recent literature emphasizes the need for architectures that balance segmentation accuracy with computational efficiency, stability, and generalization capability.

The growing use of total body imaging systems has created new opportunities for automated analysis of melanocytic nevi. By providing standardized, high-resolution images of the entire skin surface acquired over multiple examinations, these systems support longitudinal lesion monitoring and generate large datasets that require reliable methods for lesion detection, segmentation, and tracking.

Early works introduced reliable lesion localization schemes based on total-body photography, establishing the foundations

for automated lesion tracking[28]. Subsequent research focused on improving correspondence of lesions across longitudinal examinations, including methods based on three-dimensional body models for detecting new or evolving melanocytic lesions[29] and dedicated total-body scanning systems enabling automated change detection in multiple pigmented lesions[30]. Further developments addressed robustness to variations in patient pose and imaging conditions, improving lesion matching accuracy in sequential acquisitions[31].

The increasing availability of total body image datasets has stimulated the development of computer-aided diagnostic methods for pigmented skin lesions. Approaches have demonstrated automatic classification of suspicious lesions directly from wide-field images[32], as well as high diagnostic performance comparable to dermatologists using deep learning models applied to total body photographs[33]. Moreover, dedicated algorithms for total body image analysis have been proposed, highlighting challenges related to lesion scale variability, complex background structures, and limited image resolution[34].

Following studies have increasingly incorporated three-dimensional information to improve geometric consistency and measurement accuracy. Methods based on 3D body models have enabled orthorectification of lesion images, allowing more precise estimation of lesion morphology[35]. Hybrid approaches combining 3D total-body photography with convolutional neural networks have been used for reproducible automated nevus counting[36], while dedicated frameworks for lesion detection further advanced longitudinal tracking of skin changes[37]. Prospective studies have confirmed the clinical importance of total body imaging in melanoma surveillance[38].

Recent research has also emphasized the integration of 3D total-body photography with artificial intelligence techniques. Standardized annotation protocols for total-body photography datasets have been proposed to support machine learning applications[39]. Multimodal imaging approaches combining 3D photography with dermoscopy and reflectance confocal microscopy have been shown to improve melanoma detection in high-risk patients[40]. Prospective and randomized clinical studies have further validated the value of 3D total-body photography for longitudinal melanoma surveillance and early detection[41,42]. In parallel, computational methods for accurate lesion correspondence across sequential scans have been developed, including large-scale frameworks for lesion tracking in 3D body data that achieve high matching accuracy and demonstrate the growing role of artificial intelligence in automated skin lesion analysis[43].

Advances in skin lesion analysis have been driven by parallel developments in both image acquisition technologies and automated image processing methods. total body imaging systems have evolved from conventional photographic documentation to sophisticated multi-camera and three-dimensional platforms capable of longitudinal lesion monitoring, while segmentation approaches have progressed from classical image processing techniques and handcrafted feature extraction to deep learning and hybrid CNN–transformer architectures. Together, these developments have enabled increasingly accurate detection, localization, and tracking of melanocytic lesions. However, the large volume, high resolution, and unique characteristics of total body skin images continue to pose significant computational challenges. Furthermore, although state-of-the-art deep learning models often achieve excellent performance, they frequently do so at the cost of increased architectural complexity and computational demands. These considerations motivate the development of lightweight and efficient segmentation models capable of robust operation on total body images acquired in real-world clinical settings.

Methods

In this study, images were acquired using a prototype *Dermo Studio* total-body scanner (Figure 1)[34,44,45], designed for comprehensive and standardized documentation of the skin surface. The system consists of a vertical mounting column with high-resolution digital cameras and depth sensors, a pair of integrated arm supports to stabilize patient posture, and a rotating platform enabling a full 360° rotation during acquisition[45]. Illumination is provided by neutral-white LED panels (4000 K) with a total output exceeding 8000 lumens, ensuring uniform lighting and consistent color representation. Five 20 MP cameras are mounted equidistantly along the vertical column (at heights from 35 to 165 cm), capturing images up to 5488×3664 pixels. Images are acquired at 45° intervals, resulting in 40 images per scanning session. Additionally, three integrated depth sensors measure body geometry[34,45]. The fusion of depth and image data allows for the reconstruction of skin lesions in a metric space to precisely assess their size, shape, and location. Despite its complexity, the system is portable and foldable for clinical utility.

For supervised machine learning, a dataset of paired images was prepared. Each pair consists of an original skin image and a corresponding binary mask, where melanocytic lesion areas are marked with white pixels (labeled as 1) and background or lesion-free skin with black pixels (labeled as 0). The masks were manually annotated by a dermatology expert. Due to the labor-intensive nature of this process, the dataset was limited to 29 high-resolution images (stored in lossless TIFF format, with masks in PNG) acquired from five independent patients. The selected images represent various skin tones, body regions, and degrees of hair coverage, including frontal, lateral, and posterior views (Figure 2).

Processing full-resolution 20 MP images is computationally impractical for deep learning models, such as U-Net, due to GPU memory constraints and prolonged training times caused by intermediate feature maps. To address this, the images were divided into patches of 512×512 pixels with a 12.5% overlap to prevent truncation of lesions near patch edges and reduce boundary artifacts[46,47]. This process yielded approximately 3500 image–mask pairs per patient. Since most human skin is free



Figure 1. Dermo Studio total body scanner applied for image acquisition



Figure 2. Selected example images from a single patient

of lesions, the resulting dataset was highly imbalanced. To mitigate this, 85% of patches containing exclusively background were removed, ensuring a more balanced representation of lesion and non-lesion samples[46].

The resulting dataset was divided into three subsets: training, validation, and test sets. It is important to note that, during the training of deep neural networks, the validation set plays a crucial role in performance assessment and is often used to guide the selection of the loss function, optimization strategy, and the stopping point of the training process (e.g., early stopping). Consequently, the validation set cannot serve as an unbiased benchmark for the final evaluation of the model. To ensure objective assessment, data from one patient (a frontal image of the upper body) were entirely excluded and used as an independent test set. The remaining data were split into training and validation sets in an 80/20 ratio, yielding 745 and 187 samples, respectively.

The preprocessed data were then divided into training, validation, and test sets. To ensure an unbiased final performance evaluation, data from one patient (a frontal view of the upper body) were entirely excluded to serve as an independent test set. The remaining patches were split into training and validation sets in an 80/20 ratio, yielding 745 and 187 samples, respectively.

In summary, the dataset was constructed from high-resolution total body images acquired via a purpose-built scanner and paired with expert annotations. The subsequent preprocessing – including patch-based decomposition and class imbalance reduction – enabled efficient and stable model training. Due to the unique acquisition geometry and imaging conditions characteristic of this device, standard deep learning models require dedicated adaptation or retraining to achieve optimal performance on this data.

The proposed segmentation model is a modified, U-Net-like architecture designed to balance accuracy and computational

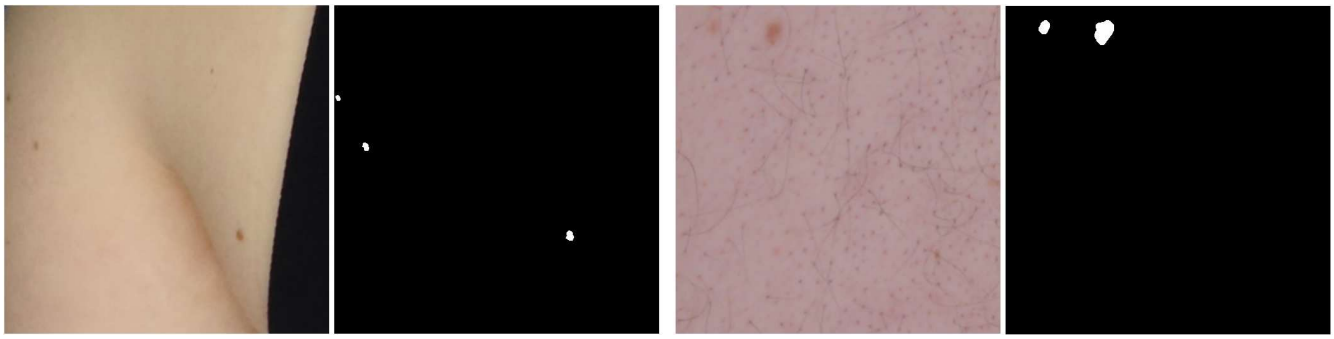


Figure 3. Example patches with corresponding masks

efficiency. While the original U-Net[11] is lightweight but lacks representational capacity in challenging conditions (low contrast, complex backgrounds)[8], advanced architectures like ResUNet++[17] achieve high accuracy at the cost of heavy computational complexity. The proposed approach bridges this gap by integrating selected enhancements inspired by ResUNet++ while maintaining a streamlined structure closer to the baseline U-Net.

In the proposed solution, the encoder consists of four levels with the number of filters increasing exponentially from 16 to 128. At each level, feature extraction is performed by two 3×3 convolutional layers followed by batch normalization[48] and the Mish activation function[49]. This improves gradient propagation and feature representation compared to the ReLU activations in U-Net[11]. To minimize architectural complexity, residual connections are omitted. Spatial resolution is progressively reduced using 2×2 max pooling, unlike in ResUNet++, which uses strided convolutions. Each encoder stage, as well as the bottleneck layer, is followed by a SE module[50] with a reduction ratio of 16 to perform channel-wise attention. Applying the SE module after the bottleneck allows for channel recalibration of the most abstract features prior to decoding – an extension not present in ResUNet++.

The bottleneck layer consists of two convolutional blocks operating on 256 channels with Mish activation[49]. Unlike ResUNet++[17], no Atrous Spatial Pyramid Pooling (ASPP) module is employed. ASPP operates by applying multiple parallel dilated (atrous) convolutions with different dilation rates, enabling the network to capture contextual information at multiple spatial scales without reducing resolution[51]. This enhances the ability to detect structures of varying size and improves segmentation of objects with complex boundaries. However, although ASPP improves multi-scale context aggregation, it significantly increases computational cost; therefore, it was omitted to maintain efficiency.

The decoder reconstructs spatial resolution using transposed convolutions, consistent with U-Net[11] and contrasting with the bilinear upsampling in ResUNet++[17], which applies bilinear upsampling. Prior to each skip connection, Attention Gates[15] are applied to filter encoder features and suppress background noise, which improves localization compared to the direct concatenation used in the original U-Net. The network output is generated by a 1×1 convolution followed by a sigmoid activation, producing a pixel-wise probability map. No additional ASPP module is applied at the output stage, further reducing inference time.

The model was trained using the AdamW optimizer[52] with a learning rate scheduler. A composite loss function was employed, combining Focal Tversky Loss[16] and Binary Cross Entropy (BCE) as a weighted sum, with $\lambda = 0.3$:

$$L = (1-\lambda)\, L_{FT} + \lambda\, L_{BCE} \tag{1}$$

The Focal Tversky Loss is based on the Tversky index (T), a generalization of the Dice-Sørensen metric, defined as:

$$T = \frac{TP + \varepsilon}{TP + \alpha\ FP + \beta\ FN + \varepsilon} \tag{2}$$

where parameters $\alpha = 0.8$ and $\beta = 0.2$ were applied to heavily penalize false negatives – a design choice motivated by the strong class imbalance, where skin lesions occupy a small fraction of the image area. The loss function takes the following form:

$$L_{FT} = (1-T)^{\gamma} \tag{3}$$

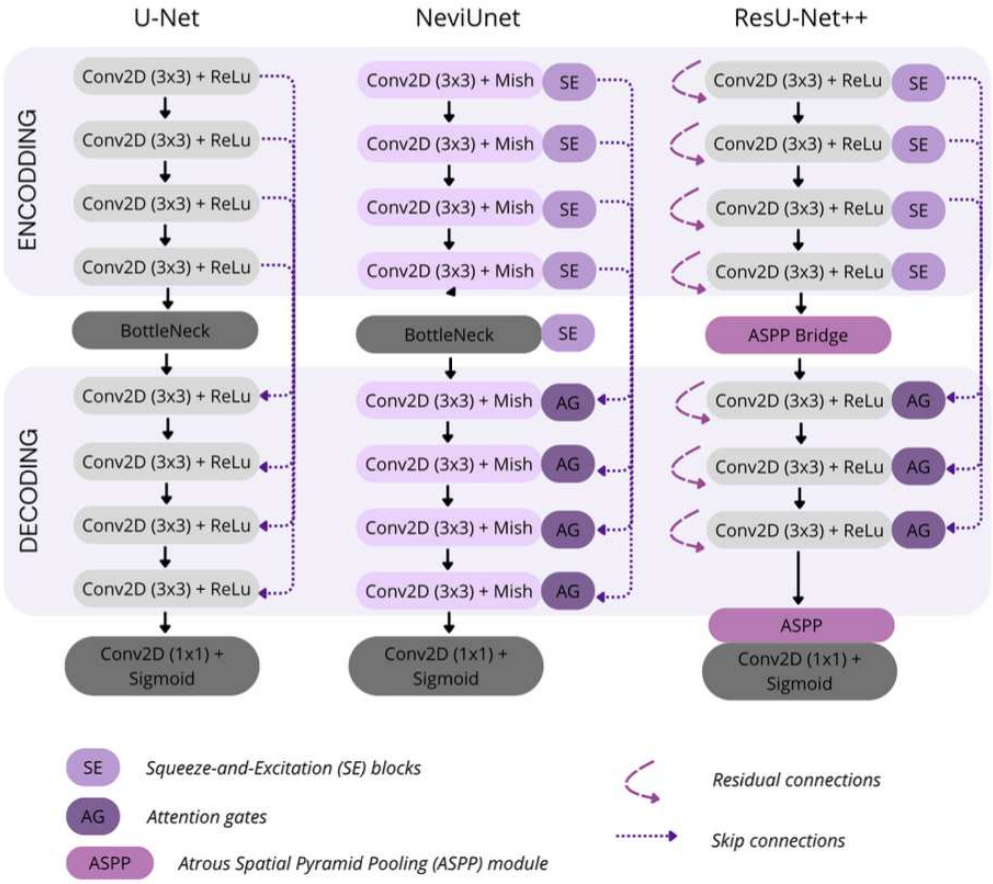


Figure 4. Comparison of the proposed solution with U-Net and ResUNet++ architectures

The parameter γ focuses the optimization on pixels with low classification confidence, thereby improving the segmentation of boundary areas and small lesions[16]. The BCE component serves as a stabilizing term during the early stages of training[53]. Training was conducted for up to 200 epochs with early stopping applied when no improvement in validation Dice-Sørensen score was observed for 20 consecutive epochs.

Overall, the proposed architecture introduces targeted modifications - namely Mish activation[49], channel attention via SE modules[50] (including after the bottleneck), spatial attention via Attention Gates[15], and a composite loss function[16,53] tailored to class imbalance – that improve segmentation performance compared to the baseline U-Net[11]. At the same time, it avoids computationally intensive components such as residual connections and ASPP modules used in ResUNet++[17], resulting in a more efficient model suitable for high-resolution total body image analysis.

Segmentation quality was evaluated on the independent test set by comparing binary prediction masks (using a 0.5 decision threshold) with reference masks at a pixel level. Predictions were categorized into true positives (*TP*), false positives (*FP*), false negatives (*FN*), and true negatives (*TN*). Due to the severe class imbalance, the Dice-Sørensen Coefficient (*DSC*) was adopted as the primary evaluation metric[54,55].

$$DSC = \frac{2\,TP}{2\,TP + FP + FN} \tag{4}$$

This metric assesses the overlap between the predicted and reference regions by assigning double weight to the intersection. To complement this quantitative analysis, a qualitative visual assessment was performed using prediction mask overlays on the original images. Different pixel classification categories were color-coded to enable a spatial evaluation of the model's errors.

Results

Number of Filters Comparison

The successful implementation is characterized by achieving high predictive accuracy while maintaining sufficient computational efficiency and model compactness to enable integration into medical devices[56]. A lightweight architecture characterized by a minimal number of parameters, low computational cost, and fast inference time is expected to be computationally efficient and suitable for deployment in resource-constrained or real-time clinical environments[57]. Motivated by the need to balance predictive accuracy with efficiency and model compactness, an experiment was conducted to empirically evaluate the optimal filter configuration across the three considered architectures.

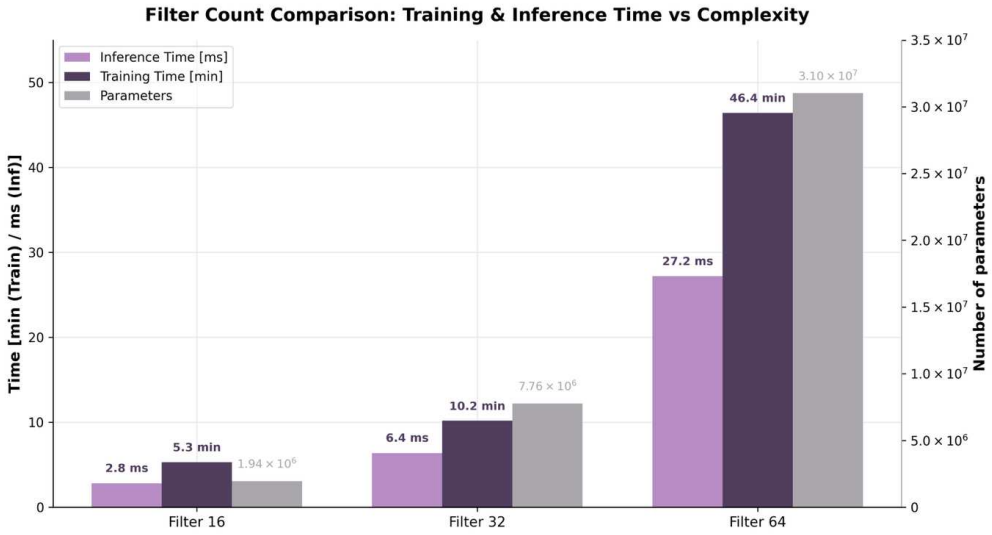


Figure 5. Training time (min), inference time (ms), and the number of trainable parameters for U-Net model with base filter counts of 16, 32, and 64

The number of convolutional filters at each encoder level directly determines the total number of trainable parameters and thus the computational cost of training and inference[58]. To quantify this relationship, the baseline U-Net architecture was evaluated with three filter configurations, defined by the base number of filters in the first encoder layer (base_f), set to 16, 32, and 64, with subsequent levels doubling this number. The results are presented in Figure 5.

Increasing the number of filters from 16 to 32 and 64 resulted in a substantial increase in model complexity, with the parameter count rising from 1.94×10^6 to 7.76×10^6 and 31.03×10^6, respectively. This was accompanied by an increase in training time (5.3 min, 10.2 min, and 46.4 min for 50 epochs) and a nearly tenfold increase in inference time (2.8 ms, 6.4 ms, and 27.2 ms). Despite this increased computational cost, all three configurations yielded similarly unsatisfactory segmentation performance, with the Dice–Sørensen coefficient close to zero, irrespective of the applied loss function – both BCE[53] and under the composite Focal Tversky Loss[16]. Consequently, the original U-Net architecture was deemed incapable of producing correct segmentation results, regardless of the number of filters used.

Efficiency and Performance Tests

To comprehensively evaluate the proposed NeviUnet architecture, an analysis was performed focusing on both computational efficiency and skin lesion segmentation performance. All measurements related to training time, inference stability and performance were conducted on a unified workstation equipped with an AMD Ryzen 7 7800X3D (8-core) processor, 64 GB of RAM, and an NVIDIA GeForce RTX 5070 GPU.

In terms of structural complexity and resource utilization (Figure 6), the baseline U-Net exhibits the lowest parameter count (1.94×10^6) and the lowest average GPU utilization (60.0%). In contrast, ResUNet++ represents the most computationally demanding architecture, requiring 4.06×10^6 parameters and achieving the highest average GPU utilization (92.6%), along with peak VRAM consumption of 12,169 MB. The proposed NeviUNet achieves a balanced trade-off, maintaining a lightweight structure comparable to the classical U-Net (1.98×10^6 parameters) while providing improved resource utilization efficiency (peak VRAM: 12,048 MB; average GPU utilization: 70.2%).

These structural characteristics directly translated into computational throughput during the training phase. ResUNet++ exhibited the highest computational cost, requiring 83 minutes for 51 epochs, corresponding to approximately 97.6 seconds

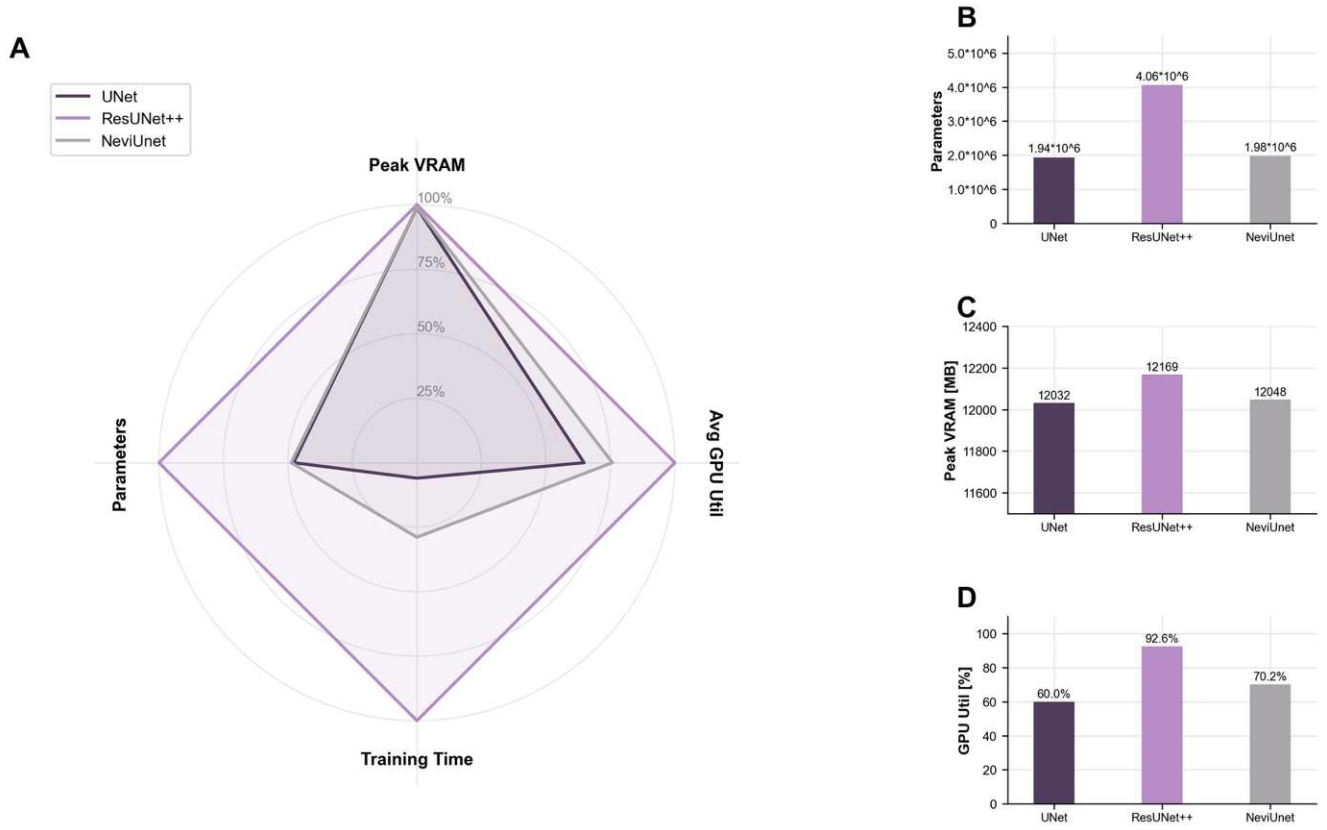


Figure 6. Hardware resource utilization: (A) radar chart of the overall hardware profile, (B) total count of parameters, (C) peak VRAM consumption, (D) average GPU utilization

per epoch. In contrast, the baseline U-Net completed training in 5 minutes over 79 epochs, with an average of 3.8 seconds per epoch; however, the process was prematurely terminated by early stopping due to stagnation in learning. The proposed NeviUNet completed the full training process in 24 minutes over 78 epochs, achieving an efficient rate of 18.5 seconds per epoch, thereby effectively combining the training speed of the baseline U-Net with enhanced feature representation capability.

Inference latency and stability analyses (Figure 7) demonstrate the operational advantage of NeviUNet over ResUNet++, particularly with increasing batch size[48]. At a batch size of 8, the inference latency of ResUNet++ increases substantially, exceeding 200 ms and accompanied by a pronounced reduction in stability (standard deviation above 40 ms). Under the same conditions, NeviUNet maintains an inference latency of approximately 32 ms with minimal temporal variance, achieving performance comparable to the baseline U-Net.

The convergence profiles of the training loss and the DSC (Figure 8) explicitly highlight the advantages of integrating the proposed architectural modifications alongside the composite loss function. Benchmarking experiments demonstrated that applying the composite loss function alone to the classical U-Net structure alters training dynamics but fails to mitigate the network's limited representational capacity, leading to late-stage instability. In contrast, the proposed model exhibits a stable, dynamic loss decay and rapid convergence, securing the highest and most consistent DSC values with minimal standard deviation across individual training runs.

Model	Loss function	Dice Sørensen Score
NeviUnet	Composite	0.729 ± 0.170
ResUNet++	Composite	0.531 ± 0.213
ResUNet++	BCE	0.491 ± 0.290

Table 1. Quantitative segmentation performance for the skin lesion class evaluated on the independent test dataset

The final evaluation of the segmentation quality was conducted on the independent test set. To ensure a rigorous assessment under severe class imbalance, the DSC computed exclusively for the actual skin lesion class (lesion) was adopted as the primary

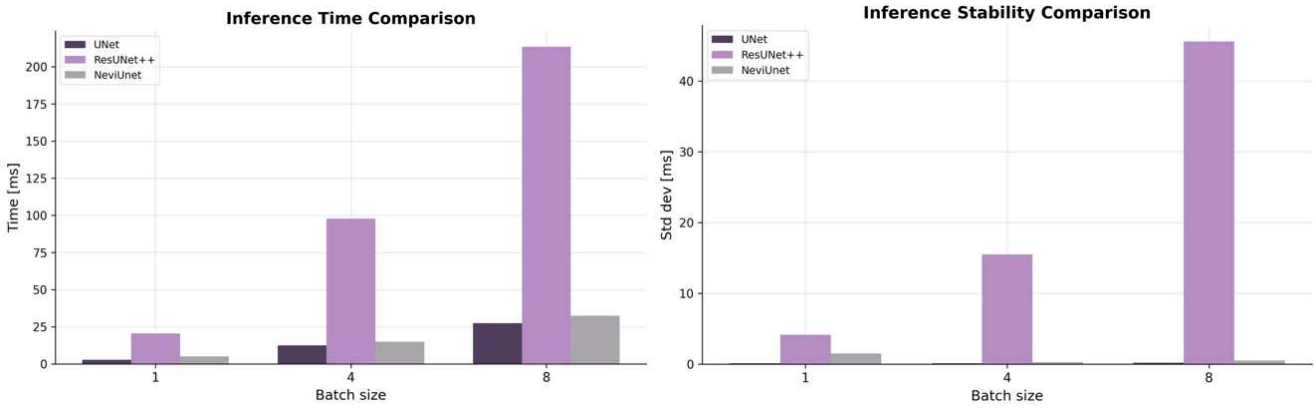


Figure 7. Mean batch prediction latency and its standard deviation, where the latter serves as a measure of inference stability

benchmark for model comparison (Table 1).

The results obtained for the classical U-Net architecture – both in its baseline configuration and in the variant employing the composite loss function – were intentionally excluded from the final summary table. In both cases, the model achieved a DSC of approximately zero, indicating a complete failure to detect lesions due to a persistent bias toward predicting all pixels as background. Analysis of the ResUNet++ models indicates that the incorporation of the composite loss function improves predictive performance. The baseline ResUNet++ trained with standard BCE loss achieved a DSC of 0.491, which increased to 0.531 with the composite loss, while the standard deviation decreased from 0.290 to 0.213, indicating improved stability.

The proposed NeviUNet achieved the best overall performance in the benchmark, with a mean DSC of 0.729. This represents an improvement of nearly 20 percentage points over the best-performing ResUNet++ variant. The corresponding violin plot (Figure 9) shows that the performance distribution of NeviUNet is strongly shifted toward higher values and exhibits a markedly reduced dispersion, confirming the robustness of the proposed architecture under challenging imaging conditions and complex geometries associated with the custom scanning system.

Discussion

The experiments performed using the baseline U-Net architecture demonstrated that simply increasing the number of convolutional filters in the first layer from 16 to 64 does not lead to satisfactory segmentation of melanocytic skin nevi (Figure 5). Even the largest evaluated model failed to produce clinically useful predictions, while both the training and inference times increased proportionally with network complexity and the number of trainable parameters. These results indicate that improved segmentation performance cannot be achieved merely by increasing model capacity; instead, the network architecture must be specifically adapted to the characteristics of the target application.

Practical deployment of automated total body skin analysis also imposes strict computational constraints. Since acquisition of a complete body scan requires approximately one minute, segmentation of all image patches should be completed within a comparable time. For the approximately 3,500 image patches generated from a single examination, this corresponds to a target inference time of about 17–20 ms per patch. The baseline U-Net with 64 initial filters exceeded this limit, requiring approximately 27 ms per patch.

ResUNet++ represents the most computationally demanding architecture evaluated in this study. Even with only 16 filters in the first encoder stage, the network contains more than four million trainable parameters and requires substantial GPU memory (Figure 6). Configurations with a larger number of filters exceeded the hardware resources available in the target system, while the use of more powerful GPUs would considerably increase system cost. Furthermore, the inference time of approximately 25 ms per patch obtained for the smallest ResUNet++ configuration is already above the maximum acceptable value for real-time operation.

In contrast, the proposed NeviUNet architecture requires only slightly more GPU resources than the baseline U-Net while maintaining fewer than two million trainable parameters (Figure 6). At the same time, it achieves an inference time below 5 ms per image patch, enabling real-time segmentation of total body images during their acquisition by the Dermo Studio scanner. These results demonstrate that the proposed architecture provides a computational efficiency, making it well suited for integration into automated total body skin imaging systems.

Moreover, the proposed architecture, NeviUNet, achieves superior segmentation performance compared with the reference models, attaining a mean DSC of 0.73 (Table 1) while maintaining low computational overhead. This result indicates that the

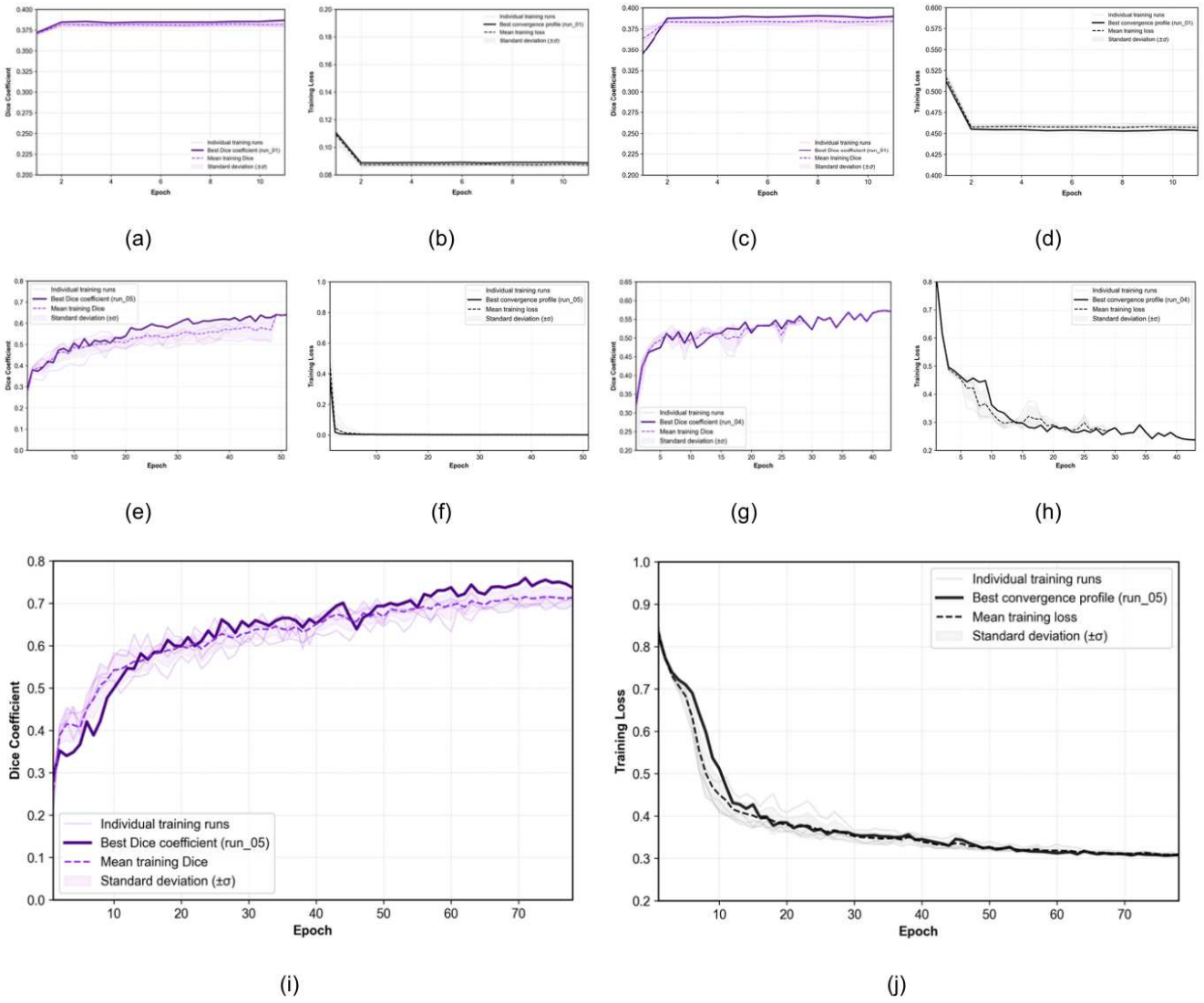


Figure 8. Training curves of the Dice Similarity Coefficient (a, c, e, g, i) and loss (b, d, f, h, j) for the evaluated architectures: (a, b) baseline U-Net, (c, d) U-Net with the composite loss function, (e, f) baseline ResUNet++, (g, h) ResUNet++ with the composite loss function, and (i, j) the proposed NeviUNet with the composite loss function

integration of Mish activation, SE modules, attention gates, and a composite loss function effectively enables reliable lesion detection under the challenging imaging conditions associated with total body photography.

In contrast, the baseline U-Net fails to detect skin lesions irrespective of the applied loss function, highlighting the limitations of architectures with insufficient representational capacity when confronted with severe class imbalance and high image complexity. When trained with standard binary cross-entropy loss, the model consistently predicts background for all pixels, while the use of Focal Tversky Loss alters the training dynamics without improving segmentation performance. This observation demonstrates that loss function modification alone cannot compensate for inherent structural deficiencies, and that stable loss convergence does not necessarily translate into meaningful segmentation outcomes.

A similar pattern is observed for ResUNet++, which, despite employing over 4 million parameters, achieves only moderate performance, with DSC of 0.49 with BCE and 0.53 with composite loss functions (Table 1). The relatively high inter-run variance further indicates limited reproducibility, likely resulting from overfitting associated with the network's structural complexity, including residual connections and ASPP modules. In comparison, NeviUNet represents a more practical solution, completing training in 24 minutes at an average GPU utilization of 70%, whereas ResUNet++ requires 83 minutes and operates near hardware limits. Importantly, NeviUNet achieves inference times of approximately 32 ms (for 8 patches batch) while maintaining stable latency, in contrast to the substantial fluctuations observed for ResUNet++, which may exceed 200 ms

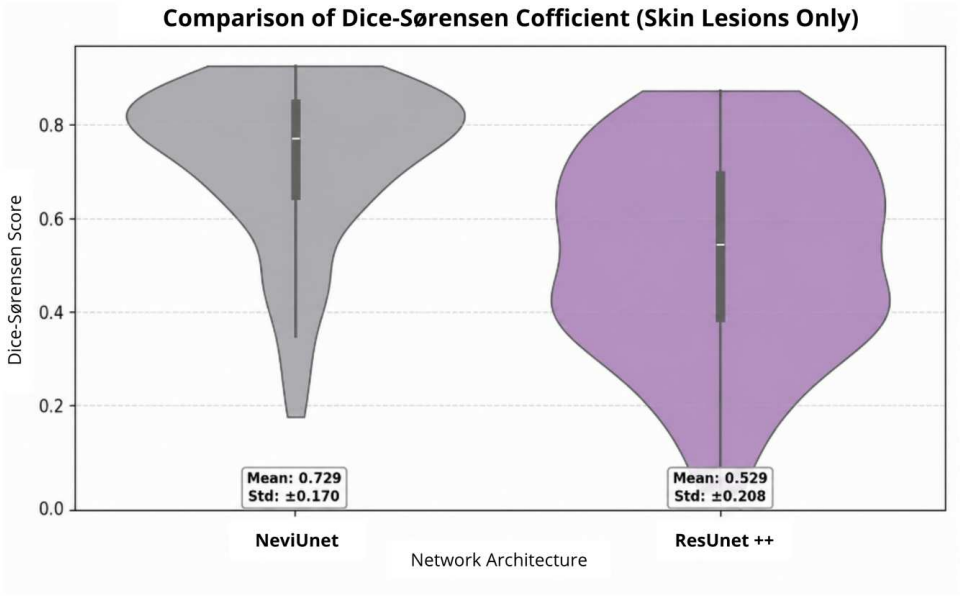


Figure 9. Violin plot comparing the DSC distributions between NeviUNet and ResUNet++

(Figure 7). These findings justify the deliberate omission of residual and ASPP components in favor of a lightweight and stable design.

The improved precision of NeviUNet is also reflected in more accurate delineation of lesion boundaries, resulting in a reduced number of false positives. Although ResUNet++ attains relatively high sensitivity, it exhibits a tendency to over-segment (Figure 10), particularly in regions with dense hair, which negatively affects the DSC due to the balanced penalization of false positives and false negatives.

These results should be interpreted in light of certain limitations, most notably the small dataset comprising a limited number of patients. Standard data augmentation was intentionally omitted, as simple geometric transformations (e.g., rotations or flips) do not introduce meaningful biological variability in total body imagery. Consequently, future work will focus on expanding the dataset to include a larger patient cohort and on developing domain-specific augmentation strategies tailored to total body dermatological imaging.

This study demonstrates that improving segmentation performance in application-specific medical imaging tasks does not necessarily require increasing the number of trainable parameters or adopting increasingly complex network architectures. Instead, superior performance can be achieved by carefully adapting and combining selected architectural components to address the characteristics and constraints of the target application. The proposed NeviUNet exemplifies this design philosophy by integrating lightweight attention mechanisms and optimized convolutional blocks into a computationally efficient U-Net-based framework. The resulting architecture provides an effective balance between segmentation accuracy, inference speed, and hardware requirements, making it suitable for practical deployment in real-time total body skin imaging systems such as the Dermo Studio scanner.

References

1. Goldstein, A. C. & Tucker, M. A. Dysplastic nevi and melanoma. *Cancer Epidemiol. Biomarkers & Prev.* **22**, 528–532, DOI: 10.1158/1055-9965.EPI-12-1346 (2013).

2. Pogorzelska-Antkowiak, A., Slowinska, M., Farinazzo, E., Conforti, C. & Agozzino, M. Spitz nevus and melanoma: evaluation with dermoscopy and reflectance confocal microscopy. *Adv. Dermatol. Allergol.* **40**, 1–5, DOI: 10.5114/ada.2023.125968 (2023).

3. Madan, V., Lear, J. T. & Szeimies, R.-M. Non-melanoma skin cancer. *The Lancet* **375**, 673–685, DOI: 10.1016/S0140-6736(09)61196-X (2010).

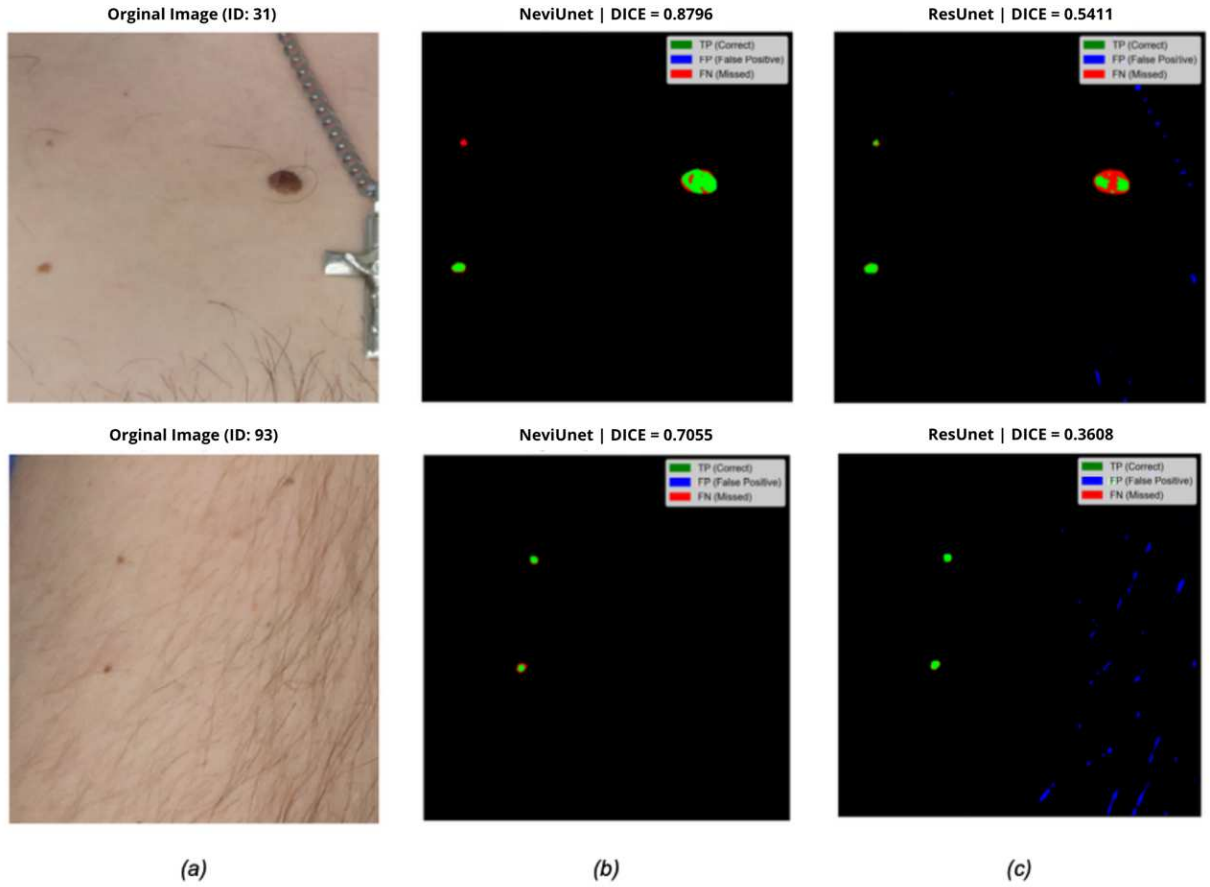


Figure 10. Examples of segmentation results compared with the ground-truth annotations: (a) input images containing skin lesions, (b) predictions generated by NeviUNet, and (c) predictions generated by ResUNet++

4. Cheng, T. W., Ahern, M. C. & Giubellino, A. The spectrum of spitz melanocytic lesions: from morphologic diagnosis to molecular classification. *Front. Oncol.* **12**, 889223, DOI: 10.3389/fonc.2022.889223 (2022).

5. Arnold, M. *et al.* Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatol.* **158**, 495–503, DOI: 10.1001/jamadermatol.2022.0160 (2022).

6. Liu, L., Tsui, Y. Y. & Mandal, M. Skin lesion segmentation using deep learning with auxiliary task. *J. Imaging* **7**, 67, DOI: 10.3390/jimaging7040067 (2021).

7. Jurakić Tončić, R. *et al.* The role of digital dermoscopy and follow-up in the detection of amelanotic/hypomelanotic melanoma in a group of high-risk patients—is it useful? *Life* **14**, 1200, DOI: 10.3390/life14091200 (2024).

8. Kassem, M. A., Hosny, K. M., Damaševičius, R. & Eltoukhy, M. M. Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics* **11**, 1390, DOI: 10.3390/diagnostics11081390 (2021).

9. Celebi, M. E. *et al.* A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **31**, 362–373, DOI: 10.1016/j.compmedimag.2007.01.003 (2007).

10. Bukhari, M. *et al.* A novel framework for melanoma lesion segmentation using multiparallel depthwise separable and dilated convolutions with swish activations. *J. Healthc. Eng.* **2023**, 1847115, DOI: 10.1155/2023/1847115 (2023).

11. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241, DOI: 10.1007/978-3-319-24574-4_28 (2015).

12. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual U-Net. *IEEE Geosci. Remote. Sens. Lett.* **15**, 749–753, DOI: 10.1109/LGRS.2018.2802944 (2018).

13. Ibtehaz, N. & Rahman, M. S. Multiresunet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* **121**, 74–87, DOI: 10.1016/j.neunet.2019.08.025 (2020).

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)*, 801–818, DOI: 10.1007/978-3-030-01234-2_49 (2018).

15. Oktay, O. *et al.* Attention U-Net: Learning where to look for the pancreas. In *Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL)* (2018).

16. Abraham, N. & Khan, N. M. A novel focal Tversky loss function with improved Attention U-Net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, 683–687, DOI: 10.1109/ISBI.2019.8759329 (2019).

17. Jha, D. *et al.* Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–230, DOI: 10.1109/ISM46123.2019.00049 (2019).

18. Mirikharaji, Z. *et al.* A survey on deep learning for skin lesion segmentation. *Med. Image Analysis* **88**, 102863, DOI: 10.1016/j.media.2023.102863 (2023).

19. Lee, T., Ng, V., Gallagher, R., Coldman, A. & McLean, D. Dullrazor: A software approach to hair removal from images. *Comput. Biol. Medicine* **27**, 533–543, DOI: 10.1016/S0010-4825(97)00020-6 (1997).

20. Kiani, K. & Sharafat, A. R. E-shaver: An improved dullrazor for digitally removing dark and light-colored hairs in dermoscopic images. *Comput. Biol. Medicine* **41**, 139–145, DOI: 10.1016/j.compbiomed.2011.01.003 (2011).

21. Kasmi, R. *et al.* Sharprazor: Automatic removal of hair and ruler marks from dermoscopy images. *Ski. Res. Technol.* **29**, e13203, DOI: 10.1111/srt.13203 (2023).

22. Bardou, D., Bouaziz, H., Lv, L. & Zhang, T. Hair removal in dermoscopy images using variational autoencoders. *Ski. Res. Technol.* **28**, 445–454, DOI: 10.1111/srt.13145 (2022).

23. Wang, J., Ruhaiyem, N. I. R. & Fu, P. A comprehensive review of U-Net and its variants: Advances and applications in medical image segmentation. *IET Image Process.* DOI: 10.1049/ipr2.70019 (2025).

24. Ding, Y. *et al.* Cth-net: A CNN and transformer hybrid network for skin lesion segmentation. *iScience* **27**, 109442, DOI: 10.1016/j.isci.2024.109442 (2024).

25. Debelee, T. G. Skin lesion classification and detection using machine learning techniques: A systematic review. *Diagnostics* **13**, 3147, DOI: 10.3390/diagnostics13193147 (2023).

26. Nie, T., Zhao, Y. & Yao, S. Multi-path segmentation network based on CNN and transformer for skin lesion image. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 3137–3153 (2024).

27. Iqbal, S. *et al.* Tesl-net: A transformer-enhanced CNN for accurate skin lesion segmentation. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 88–95 (2024).

28. Cho, T. S., Freeman, W. T. & Tsao, H. A reliable skin mole localization scheme. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1–8 (2007).

29. Bogo, F., Romero, J., Peserico, E. & Black, M. J. Automated detection of new or evolving melanocytic lesions using a 3D body model. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 593–600, DOI: 10.1007/978-3-319-10404-1_74 (2014).

30. Korotkov, K., Quintana, J., Puig, S., Malvehy, J. & Garcia, R. A new total body scanning system for automatic change detection in multiple pigmented skin lesions. *IEEE Transactions on Med. Imaging* **34**, 317–338, DOI: 10.1109/TMI.2014.2357715 (2015).

31. Korotkov, K. *et al.* An improved skin lesion matching scheme in total body photography. *IEEE J. Biomed. Heal. Informatics* **23**, 586–598, DOI: 10.1109/JBHI.2018.2855409 (2019).

32. Birkenfeld, J. S., Tucker-Schwartz, J. M., Soenksen, L. R., Avilés-Izquierdo, J. A. & Marti-Fuster, B. Computer-aided classification of suspicious pigmented lesions using wide-field images. *Comput. Methods Programs Biomed.* **195**, 105631, DOI: 10.1016/j.cmpb.2020.105631 (2020).

33. Soenksen, L. R. *et al.* Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci. Transl. Medicine* **13**, eabb3652, DOI: 10.1126/scitranslmed.abb3652 (2021).

34. Strzelecki, M. H. *et al.* Skin lesion detection algorithms in whole body images. *Sensors* **21**, 6639, DOI: 10.3390/s21196639 (2021).

35. Szczypiński, P. M. & Sprawka, K. Orthorectification of skin nevi images by means of a 3D model of the human body. *Sensors* **21**, 8367, DOI: 10.3390/s21248367 (2021).

36. Betz-Stablein, B. *et al.* Reproducible naevus counts using 3D total body photography and convolutional neural networks. *Dermatology* **238**, 4–11, DOI: 10.1159/000517218 (2022).

37. Zhao, M., Kawahara, J., Abhishek, K., Shamanian, S. & Hamarneh, G. Skin3d: Detection and longitudinal tracking of pigmented skin lesions in 3D total-body textured meshes. *Med. Image Analysis* **77**, 102329, DOI: 10.1016/j.media.2021.102329 (2022).

38. Koh, U. *et al.* Mind Your Moles study: Protocol of a prospective cohort study of melanocytic naevi. *BMJ Open* **8**, e025857, DOI: 10.1136/BMJOPEN-2018-025857 (2018).

39. Primiero, C. A. *et al.* A protocol for annotation of total body photography for machine learning to analyze skin phenotype and lesion classification. *Front. Medicine* **11**, 1380984, DOI: 10.3389/fmed.2024.1380984 (2024).

40. Hobelsberger, S., Steininger, J., Meier, F. E., Beissert, S. & Gellrich, F. F. Three-dimensional total body photography, digital dermoscopy, and in vivo reflectance confocal microscopy for follow-up assessments of high-risk patients for melanoma: A prospective, controlled study. *Dermatology* **240**, 803–813, DOI: 10.1159/000541894 (2024).

41. Soyer, H. P. *et al.* 3d total-body photography in patients at high risk for melanoma. *JAMA Dermatol.* **161**, 472, DOI: 10.1001/jamadermatol.2025.0211 (2025).

42. Serra-García, L. *et al.* Clinical outcomes of 3d-total body photography and digital dermoscopy for surveillance of high-risk melanoma patients: a prospective longitudinal observational study. *Eur. J. Cancer* **231**, 116053, DOI: 10.1016/j.ejca.2025.116053 (2025).

43. Huang, W.-L. *et al.* Revisiting lesion tracking in 3d total body photography. *Med. Image Analysis* **102**, 103963 (2026).

44. Strzelecki, M. *et al.* Artificial intelligence in the detection of skin cancer: state of the art. *Clin. Dermatol.* **42**, 280–295, DOI: 10.1016/j.clindermatol.2023.12.022 (2024).

45. Skopia Estetic Clinic Sp. z o.o. Dermo studio — automatyczne mapowanie ciała. https://dermo-studio.eu (2026). Accessed May 21, 2026.

46. Cira, C. I., Manso-Callejo, M. A., Yokoya, N., Sălăgean, T. & Badea, A. C. Impact of tile size and tile overlap on the prediction performance of convolutional neural networks trained for road classification. *Remote. Sens.* **16**, 2818, DOI: 10.3390/rs16152818 (2024).

47. Cira, C. I., Manso-Callejo, M. A., Alcarria, R., Iturrioz, T. & Arranz-Justel, J. J. Insights into the effects of tile size and tile overlap levels on semantic segmentation models trained for road surface area extraction from aerial orthophotography. *Remote. Sens.* **16**, 2954, DOI: 10.3390/rs16162954 (2024).

48. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 448–456 (2015).

49. Misra, D. Mish: a self regularized non-monotonic activation function. In *Proceedings of the British Machine Vision Conference (BMVC)*, DOI: 10.5244/c.34.191 (2020).

50. Ni, J. C., Lee, S. H., Shen, Y. C. & Yang, C. S. Improved u-net based on resnet and se-net with dual attention mechanism for glottis semantic segmentation. *Med. Eng. & Phys.* **136**, 104298, DOI: 10.1016/j.medengphy.2025.104298 (2025).

51. Wang, J. *et al.* Dds-deeplabv3+: a lightweight deformable convolutional network for cloud detection in remote sensing imagery. *Remote. Sens.* **18**, 621, DOI: 10.3390/rs18040621 (2026).

52. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).

53. Yeung, M., Sala, E., Schönlieb, C.-B. & Rundo, L. Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* **95**, 102026, DOI: 10.1016/j.compmedimag.2021.102026 (2022).

54. Taha, A. A. & Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29, DOI: 10.1186/s12880-015-0068-x (2015).

55. Maier-Hein, L. *et al.* Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217, DOI: 10.1038/s41467-018-07619-7 (2018).

56. Dinh, B. D., Nguyen, T. T., Tran, T. T. & Pham, V. T. 1m parameters are enough? a lightweight cnn-based model for medical image segmentation. In *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1279–1284 (2023).

57. Gulzar, Y., Ya'u, B. I., Alkanan, M. & Onn, C. W. Scnet: a lightweight cnn with depthwise and se modules for skin lesion classification. *Comput. Methods Biomech. Biomed. Eng. Imaging & Vis.* **13**, 2576198, DOI: 10.1080/21681163.2025.2576198 (2025).

58. Li, H., Kadav, A., Durdanovic, I., Samet, H. & Graf, H. P. Pruning filters for efficient convnets. In *International Conference on Learning Representations (ICLR)* (2017).

Acknowledgments

Dataset

The image dataset used in this study was provided by Skopia Estetic Clinic Sp. z o.o., headquartered in Kraków, and was acquired within the framework of the project POIR.04.01.04-00-0125/18, *Development of a device to support early stage examination of skin lesions, including melanoma using computer imaging, spatial modeling, comparative analysis and classification methods*, funded by the European Regional Development Fund and the National Centre for Research and Development. The dataset was labeled and curated by Dr. Dorota Katarzyna Wielowieyska-Szybińska. In the presented study, we were able to use the dataset due to our long-standing collaboration with Skopia Estetic Clinic under a licensing agreement between Skopia Estetic Clinic and Lodz University of Technology.

The dataset was acquired in accordance with the principles of the Declaration of Helsinki. Ethical review and approval were waived by the Ethics Committee of the Regional Medical Chamber in Kraków due to the fully non-invasive nature of the study and the measures taken to ensure the anonymity of all participating volunteers. Informed consent was obtained from all subjects involved in the study. The data presented in this study are not publicly available because optical total body images and 3D scans of patients constitute sensitive medical information.

The data that support the findings of this study are available from Skopia Estetic Clinic; however, restrictions apply to the availability of these data, which were used under license for the current study and are therefore not publicly available. Access to the data may be granted by Skopia Estetic Clinic upon reasonable request and under conditions determined by the data owner. The authors are not authorized to provide the data directly but may assist interested researchers in contacting Skopia Estetic Clinic.

AI-Assisted technologies

During the preparation of this work, the authors used ChatGPT to translate sections of the text from Polish and to assist with style and grammar corrections. All content generated or refined with this tool was carefully reviewed and edited by the authors, who take full responsibility for the final content of the publication.

Author contributions statement

K.D.: Investigation, Methodology, Software, Validation, Visualization, Writing; P.M.S.: Conceptualization, Methodology, Formal analysis, Supervision, Writing. All authors reviewed the manuscript.

Additional information

Competing interests

The authors declare no competing interests.

Funding

No funding was received for this manuscript.