SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 7, No. 3, 2024, 449-459
DOI: 10.35377/saucis...1543993

RESEARCH ARTICLE

An Evaluation of Skin Lesion Segmentation Using Deep Learning Architectures

Gökçen Çetinel[1], Bekir Murat Aydın[1], Sevda Gül[2*], Devrim Akgün[3], Rabia Öztaş Kara[4]

[1]Department of Electrical and Electronics Engineering, Faculty of Engineering, Sakarya University, Sakarya, Türkiye
[2]Department of Electronics and Automation, Adapazarı Vocational School, Sakarya University, Sakarya, Türkiye
[3]Department of Software Engineering, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, Türkiye
[4]Department of Dermatology, Training and Research Hospital, Sakarya University, Sakarya, Türkiye

Corresponding author:
Sevda Gül,
Department of Electronics and Automation,
Adapazarı Vocational School,
Sakarya University, Sakarya, Türkiye,
gulsevda@sakarya.edu.tr

Article History:
Received: 05.09.2024
Accepted: 20.11.2024
Published Online: 25.12.2024

ABSTRACT
Skin lesion segmentation for recognizing and defining the boundaries of skin lesions in images is proper for automated analysis of skin lesion images, especially for the early diagnosis and detection of skin cancers. Deep learning architectures are an efficient way to implement segmentation once a skin lesion dataset is provided with ground truth images. This study evaluates deep learning architectures on a hybrid dataset, including a private dataset collected from a hospital and a public ISIC dataset. Four different test cases exist in the analysis where the combinations of public and private datasets are used as train and test datasets. Experimental results include Unet, Unet++, DeepLabV3, DeepLabV3++, and FPN segmentation architectures. According to the comparative evaluations, mixed datasets, where public and private datasets were used together, provided the best results. The evaluations also show that the collected dataset with ground truth images provided promising results.

Keywords: Skin lesion segmentation, Deep learning architectures, Unet, DeepLabV3, Feature pyramids network

1. Introduction

Skin cancer is a significant global health concern, and effective treatment depends on early and accurate identification [1]. Skin lesion segmentation is helpful for automated analysis of skin lesion images, especially for the early diagnosis and monitoring of skin cancers. Segmentation of the skin lesions process helps recognize and draw the boundaries of skin lesions in images to increase the efficiency and accuracy of skin cancer diagnosis. Accurate segmentation is essential for quantitative assessment of lesion characteristics and monitoring changes in lesions over time. This can be challenging due to variations in lesion size, shape, color, texture, and factors like lighting conditions and image quality. Skin cancer, particularly melanoma, poses a growing global health threat, with millions of new cases diagnosed each year. Early skin cancer detection, primarily through automated skin lesion segmentation, is critical in improving survival rates. Lesion segmentation enables clinicians to delineate cancerous tissues precisely, providing vital information for diagnosis and treatment planning.

Skin lesion segmentation aims to outline the lesions' boundaries accurately, separating them from the surrounding healthy skin. Many approaches and designs have been developed, each with unique strengths and uses. Segmentation techniques can be manual, where trained professionals outline the lesion boundaries manually, or automated, where computer algorithms are employed to perform the segmentation automatically. Once adequately trained, machine learning models can reach high levels of accuracy in segmentation [2]. Advances in skin lesion segmentation with deep learning approaches, a popular machine learning subclass, have significantly contributed to diagnosing skin diseases [3]. Deep learning models have improved the field of skin lesion segmentation by giving powerful tools for contrasting skin lesions from healthy skin in dermatological images [4].

This study aims to evaluate the effectiveness of deep learning architectures in diagnosing skin lesions by analyzing their performance across a combination of public International Skin Imaging Collaboration (ISIC) datasets [5]–[8] and private datasets confidential to Sakarya University Training and Research Hospital (SEAH). Experimental evaluations involve four distinct test cases where the training and testing datasets consist of various mixes of these public and private datasets, aiming to assess how well these models can learn from and adapt to the differences in data sources, improving the accuracy and reliability of skin lesion diagnosis. Experimental results include a comparative evaluation of Unet, Unet++, DeepLabV3,

Cite as: G. Çetinel, B. M. Aydın, S Gül, D. Akgün, R. Öztaş Kara, “An Evaluation of Skin Lesion Segmentation Using Deep Learning Architectures,” *Sakarya University Journal of Computer and Information Sciences,* vol.7, no. 3, pp. 449-459, 2024. Doi: 10.35377/saucis…1543993

DeepLabV3++, and Feature Pyramid Network (FPN) segmentation architectures.

2. Related Works

The progress made in the development of deep learning models has helped improve skin lesion segmentation in recent years. Researchers focused on designing better architectures to improve the accuracy of skin lesion segmentation. Ronneberger et al. introduced the Unet architecture based on Convolutional Neural Network (CNN) with a symmetric encoder-decoder structure with skip connections [9]. They developed the Unet architecture primarily for medical segmentation tasks, but it has also been found to be helpful in other disciplines [10]–[13]. Its structure is characterized by a reduction path to capture essential features and an expanding path that enables accurate localization of the segmentation region. Zhou et al. proposed Unet++, a modified version of Unet that introduces nested, dense skip pathways for improved segmentation accuracy [14]. DeepLabV3 and DeepLabV3+, developed by Chen et al., utilize atrous convolution and Atrous Spatial Pyramid Pooling (ASPP) to effectively segment objects at various scales, proving highly beneficial for skin lesion analysis [15]. Oktay et al. added attention blocks to the Unet architecture to focus on relevant features for segmentation, improving the precision of lesion identification [16]. Yi et al. structured the problem as an adversarial application using Generative Adversarial Networks (GANs) for skin lesion segmentation, which resulted in more refined segmentation outputs [17]. SegNet, proposed by Badrinarayanan et al., employs an encoder-decoder architecture for pixel-wise classification tasks. They have adapted it for skin lesion segmentation to address variations in lesion appearance effectively [18]. Fully Convolutional Networks (FCNs), introduced by Long et al., facilitate semantic segmentation by allowing end-to-end segmentation with input of any size [19]. Alom et al. developed R2Unet, which integrates Recurrent Neural Networks (RNNs) with Unet to capture image spatial dependencies better [20]. Milletari et al. proposed V-Net, a volumetric extension of Unet designed for 3D image segmentation, adapted by researchers for 3D skin lesion segmentation tasks [21]. Ibtehaz and Rahman introduced MultiResUNet, which improves Unet by combining multi-resolution analysis to better capture features at various scales [22]. Sharen et al. combined Unet and FPN for skin lesion segmentation with encoder modifications [23]. Thanks for the contributions of the reviewer.

Furthermore, recent studies have demonstrated the efficacy of Unet++ and DeepLabV3+ in various medical imaging applications. For instance, Unet++ has shown effective boundary detection for complex medical imaging tasks like breast cancer and lung nodule segmentation. It enhances segmentation precision through dense skip connections, refining feature extraction in these applications [24]. Similarly, DeepLabV3+ has outperformed traditional models in retinal image analysis and liver lesion segmentation by effectively utilizing multi-scale feature extraction [25]. Recent advancements in skin lesion segmentation have included attention mechanisms, which focus on the most relevant image regions, improving segmentation accuracy [26]. GANs have also been applied to generate high-quality segmentations by modeling lesion boundaries more precisely [27].

Additionally, transformer-based models have emerged as promising due to their ability to capture long-range dependencies [28]. Many works in the literature focus on segmentation architectures for improving skin lesion segmentation by using public datasets [29], [30]. This study used well-known Unet, Unet++, DeepLabV3, DeepLabV3++, and FPN segmentation architectures. We focused on the performances of the popular segmentation architectures over public, private, and mixed datasets.

3. Methodology

This section introduces the segmentation models investigated in the proposed study and describes the used dataset.

3.1. Dataset

The study database was obtained by combining public and private databases. A total of 4228 images were utilized in the study, comprising 3463 images acquired from public databases and 765 images sourced from private database. The comprehensive database encompasses diverse dermoscopy images corresponding to various lesion types.

The ISIC datasets are an essential public resource in dermatology for skin lesion segmentation and classification research. They comprise a vast collection of publicly accessible dermoscopic images with detailed annotations, covering various skin lesions, including melanoma. These datasets, enriched with metadata like patient demographics and clinical diagnoses, are vital for developing and validating machine-learning algorithms for automated skin lesion analysis. The ISIC datasets support challenges and competitions, such as the ISIC Challenge, promoting advancements in diagnostic accuracy and dermatological research. ISIC is an open-access database that was extended in 2016, 2017, 2018, 2019, and 2020.

This study's curated ISIC database includes 4075 images with ground truth annotations. However, there exist recurring images. So, we considered the database before performing the segmentation algorithm, and 3463 images remained. The database encompasses various diagnostic categories commonly encountered in medical and dermatological image analysis. These categories include actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions. Each of these categories represents a distinct skin condition or pathology, and including such a variety of conditions makes the database valuable for research and training machine learning models in the context of dermatological diagnosis and skin cancer detection.

The private database for the study, which consists of 765 images, was constructed by a dermatologist from Sakarya University Training and Research Hospital with the permission of the ethics committee. The collected private dataset consists of high-resolution dermoscopic images with a resolution of 1024x1024 pixels. The dermatologist manually determined the lesion area for the images in the private database, thus creating ground truth masks. The included skin lesion types were melanoma, melanocytic nevus, lentigo, seborrheic keratosis, angioma, actinic keratosis, Bowen's disease, trichilemmal cyst, keratoacanthoma, hemangioma, malignant melanoma, basal cell carcinoma, pyogenic granuloma, Spitz nevus, and Merkel cell carcinoma. However, the dataset has limitations, including a smaller sample size than public datasets like ISIC and potential biases introduced during the manual annotation. The training, validation, and test sets were assembled by integrating images from public and private databases. These databases were systematically divided into proportions of 60%, 20%, and 20%, respectively.

Furthermore, the database intentionally includes challenging lesion images to ensure a robust evaluation of the proposed segmentation process. These include lesions found on individuals with dark skin tones, lesions covering the entire imaging field, and lesions closely resembling the skin color. Figure 1 displays examples of lesion images and their corresponding ground truth masks from our private dataset, which comprises 765 images, alongside the ISIC dataset containing 3463 images.

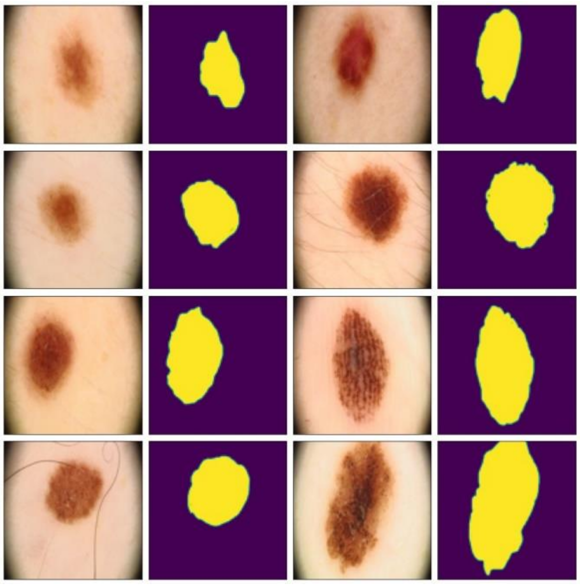


Figure 1. Example Images and Their Masks from the Private Dataset

3.2. Segmentation Models

Segmentation models mainly consist of encoder and decoder structures. The encoder filters out important features by following a narrowing path of convolution and scaling layers. The decoder, again formed by the convolution and scaling layers, obtains the output image of the same size from the features obtained. One of the architectures we used for segmentation tasks is the Unet model, which is one of the pioneer implementations with convolutional layers [9]. The Unet and its derivatives are fundamental in medical image segmentation, particularly for tasks like skin lesion segmentation. Unet's architecture and later modifications, designed for biomedical image segmentation, have proven helpful in distinguishing skin lesions from surrounding healthy skin. A variation of Unet is Unet++, which uses a series of nested, dense skip pathways that aim to reduce the semantic gap between the feature maps of the encoder and decoder parts of the network [14]. This design enhances the model's ability to capture fine-grained details and improves segmentation performance, particularly in challenging areas of skin lesion images. In Unet++, nested skip pathways introduce dense skip connections between the encoder and decoder parts of the network. These connections bridge the semantic gap between lower-level and higher-level feature maps, enhancing the model's ability to capture fine details. The architecture integrates multiple dense blocks, each designed to improve the segmentation performance by refining the feature maps at different stages. Figure 2 shows the basic structure of Unet++, where skip pathways between the encoder and decoder blocks are illustrated. In Figure 2, the Unet++ architecture shows how nested skip connections enable fine-grained segmentation by continuously refining the feature maps at different resolutions. Another model we used for comparison is the DeepLabv3 architecture, which includes decoder encoder structures like Unit. However, DeepLabV3 uses dilated convolutions and ASPP to identify the features of the backbone network. The ASPP module in DeepLabV3+ captures context at multiple scales by applying dilated convolutions with varying rates. This allows the model to handle objects of varying sizes effectively, enhancing segmentation accuracy in complex images without clearly defined object boundaries. DeepLabV3+ uses the DeepLabV3 encoder module and atrous convolution to adjust the resolution of features [15] as shown in Figure 3. DeepLabV3+'s ASPP module, which extracts contextual information from multi-scale receptive fields, improves the model's ability to handle lesions of varying sizes and is demonstrated in Figure 3. The last model we used for comparison is the FPN architecture, which is helpful for scale variations in object detection tasks [31]. The multi-scale feature pyramid enables FPNs to collect features of objects regardless of their size in the image. Figure 4 shows the basic structure of FPN, where a multi-scale feature pyramid is formed by integrating feature maps from several stages of a backbone model like ResNet. The FPN’s multi-scale feature pyramid, which allows the model to detect objects of different scales by combining low- and high-level feature maps effectively, is also highlighted in the figure.

These architectures were selected due to their proven effectiveness in medical image segmentation tasks. Unet is well-suited for biomedical segmentation because it captures fine details through skip connections between encoder and decoder layers. However, it struggles with scale variations in lesions. Unet++ improves upon Unet by introducing nested skip connections, reducing the semantic gap between feature maps at different stages. This results in improved segmentation accuracy, particularly for complex lesion boundaries. DeepLabV3 and DeepLabV3+ utilize ASPP to handle objects at multiple scales, making them highly effective for varying-sized lesions. However, their computational complexity can be a drawback in resource-constrained environments. FPN, with its multi-scale feature pyramid, excels in detecting lesions of different sizes but may lose fine details in small objects. FPN's ability to handle scale variations further supports robust segmentation, making these models ideal.

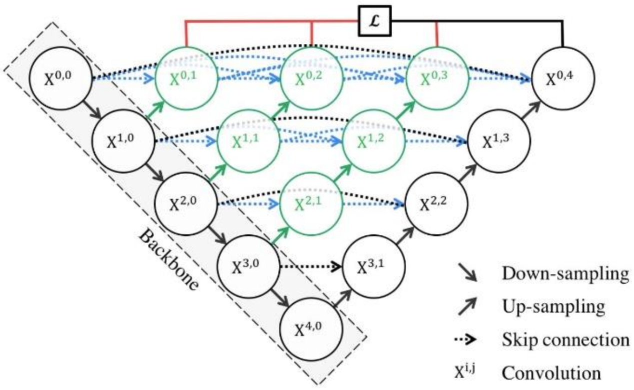


Figure 2. Unet++ Architecture [14]

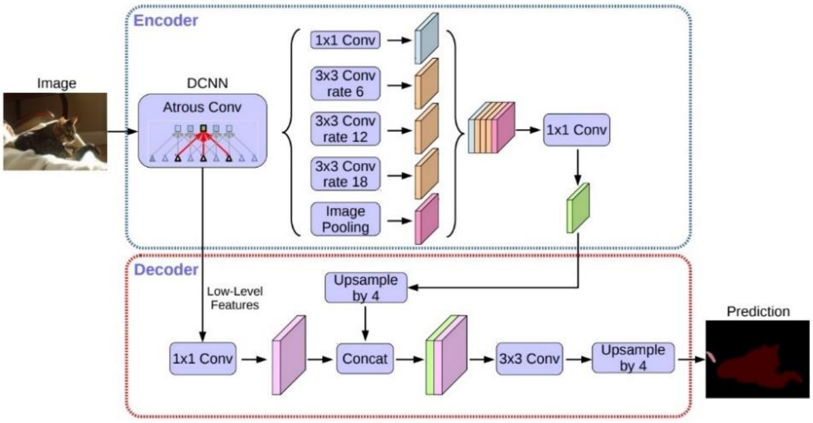


Figure 3. DeepLabV3+ Architecture [15]

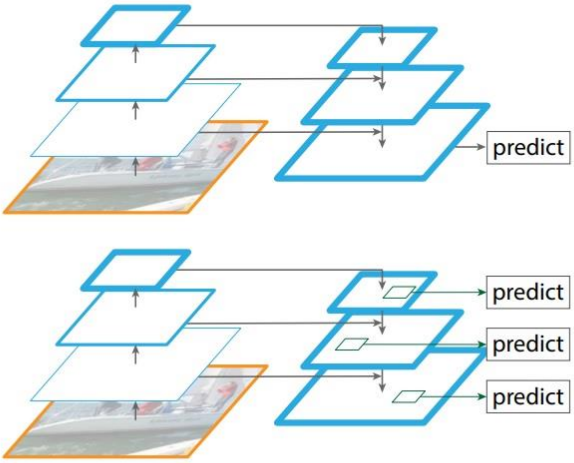


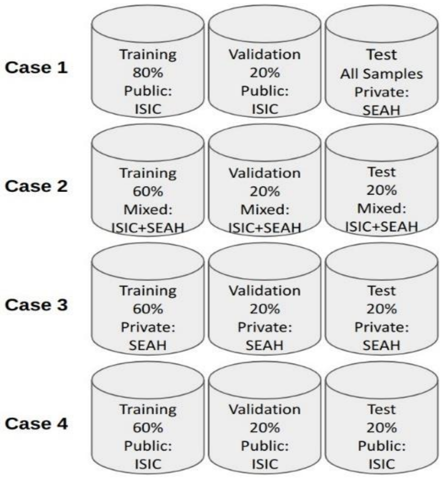
Figure 4. FPN Architecture [26]

4. Experimental Results

Experimental results were obtained using the programming language Python 3.11 and the deep learning library PyTorch 1.12.1 on the Ubuntu 20.04 operating system. The hardware for the experimental results utilizes an NVIDIA GeForce RTX 3060 12GB GPU alongside an AMD Ryzen 2700 CPU with 32GB of memory. The models were trained using the Adam optimizer with an initial learning rate 0.001. A batch size 16 was selected to balance memory usage and training efficiency. Training was conducted over 100 epochs, with early stopping employed to prevent overfitting. Data augmentation techniques such as horizontal and vertical flipping, random rotation, and brightness adjustment were applied to increase the variability of the training set and improve generalization. The training process took approximately 12 hours on an NVIDIA GeForce RTX 3060 GPU.

Four different experimental scenarios were created for training and testing segmentation models in experimental studies. The first and fourth cases involve training on the public dataset, with subsequent testing conducted on private and public datasets. Case 3 trains and tests the model on a private dataset. Conversely, Case 2 performs training and test tasks on the hybrid dataset. This approach enables a comprehensive evaluation of segmentation performance, capturing a spectrum of conditions encountered across public and private image datasets. Conducting a study focused on skin lesion segmentation, a private dataset is being curated alongside publicly available ones for training. Utilizing both public and private datasets offers several advantages. Public data aids in generalizing the model, enabling it to comprehend diverse scenarios effectively.

Conversely, the private dataset focuses on specific application domains, enhancing the model’s performance within those contexts. Employing a mixed training approach that integrates both datasets improves the model’s adaptability to general and domain-specific data. This approach promotes diversity in training and bolsters the reliability of the model’s predictions. Figure 5 presents the distribution of public and private datasets used across the four test cases. The results indicate combining public and private datasets yields the highest performance across all models.



Case 1	Training 80% Public: ISIC	Validation 20% Public: ISIC	Test All Samples Private: SEAH
Case 2	Training 60% Mixed: ISIC+SEAH	Validation 20% Mixed: ISIC+SEAH	Test 20% Mixed: ISIC+SEAH
Case 3	Training 60% Private: SEAH	Validation 20% Private: SEAH	Test 20% Private: SEAH
Case 4	Training 60% Public: ISIC	Validation 20% Public: ISIC	Test 20% Public: ISIC

Figure 5. Dataset Distribution for Four Experimental Scenarios

Figure 6 displays the segmentation results for Unet++, DeepLabV3+, and FPN. Notably, FPN achieves superior segmentation quality, particularly in challenging cases where lesions closely resemble surrounding tissue. Original lesion images, manually prepared ground truth images, and the predicted images achieved by the utilized segmentation methods are illustrated in each figure column, respectively. As can be seen from the sample segmentation results, the ground truth images are compatible with the data. The segmentation models produced similar results, and slightly better results were obtained with the FPN architecture. The detailed comparison shows FPN’s robustness in handling such variations.

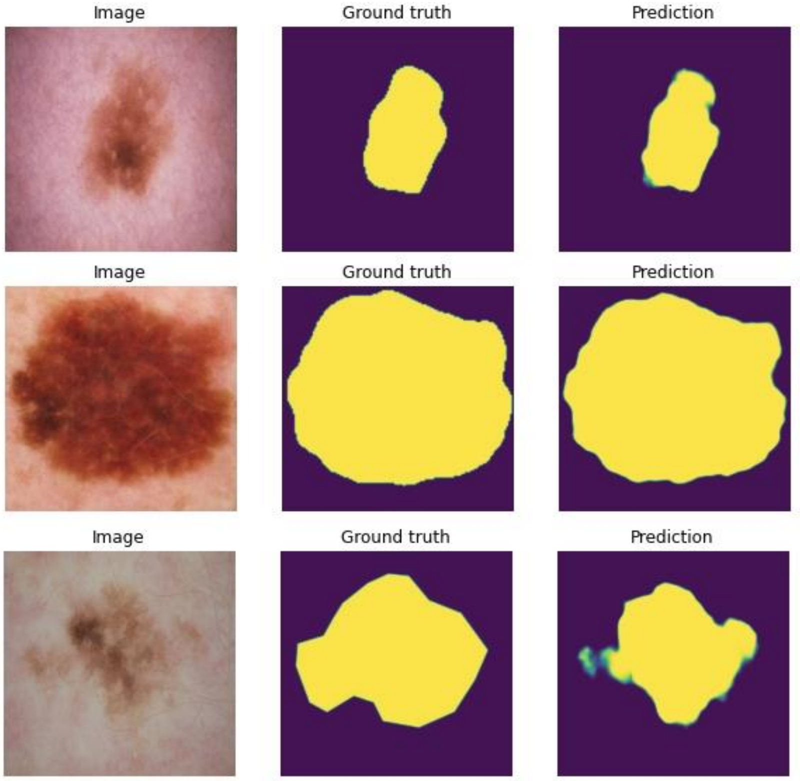


Figure 6. Sample Segmentation Results Showing Original Images, Ground Truth Masks, and Predicted Masks

Various metrics are calculated by comparing the predicted and ground truth masks to evaluate the performance segmentation methods. This study uses two standard metrics to show the models' effectiveness. The Dice coefficient and Intersection of Unıon (IoU) are two prevalent similarity measures. Dice is calculated as twice the overlap area between the predicted and ground truth masks, divided by the total area of both masks, yielding a value between 0 and 1, with 1 indicating perfect overlap. IoU or Jaccard index, on the other hand, is the ratio of the intersection area to the union area of the predicted and ground truth masks. These metrics are crucial for assessing the model's performance, especially in datasets with limited examples. The equations for Dice and IoU metrics can be given as follows:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

where $|A|$ and $|B|$ are the number of pixels in the predicted and ground truth segmentation sets, respectively. $|A \cap B|$ and $|A \cup B|$ show the intersection and union of these sets, respectively.

On the other side, the pixel-level confusion matrix, which includes True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values, provides a detailed breakdown of the models’ predictions. Accuracy, calculated as the ratio of correctly predicted pixels to the total number of pixels, offers an overall measure of the model's correctness. Sensitivity measures the model's ability to correctly capture all lesion pixels, indicating how effectively the model segments true lesion areas without missing parts. Specificity indicates the model's capacity to exclude non-lesion pixels, reducing false positive segmentation around lesions. High specificity shows reliable background separation. Together, these metrics comprehensively evaluate the model's strengths and weaknesses, enabling precise adjustments to enhance its performance. The equations of the metrics can be expressed as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP+FP}}{\text{TP+TN+FP+FN}} \quad (5)$$

In the presented study, five metrics, Dice, IoU, sensitivity, specificity, and accuracy, were calculated to provide a more comprehensive assessment. Tables 1-5 give the results of the utilized segmentation methods.

Table 1. Methods Evaluations for IoU Metric

Methods	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Average**
Unet	0.7623	0.8105	**0.8075**	0.8177	**0.7995**
DeepLabV3	**0.7624**	0.8126	0.7967	0.8140	0.7964
Unet++	0.7507	0.8099	0.8071	**0.8229**	0.7976
DeepLabV3+	0.7549	0.8177	0.7986	0.8219	0.7983
FPN	0.7508	**0.8146**	0.8028	0.8139	0.7955

Table 2. Methods Evaluations for Dice Metric

Methods	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Average**
Unet	0.8610	0.8765	0.8837	0.8842	**0.8763**
DeepLabV3	**0.8636**	0.8734	0.8871	0.8767	0.8753
Unet++	0.8519	0.8731	0.8753	**0.8859**	0.8716
DeepLabV3+	0.8587	0.8778	0.8747	0.8851	0.8741
FPN	0.8536	**0.8939**	**0.8874**	0.8797	0.8741

Table 3. Methods Evaluations for Sensitivity Metric

Methods	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Average**
Unet	0.9272	0.9346	0.9449	0.9249	0.9329
DeepLabV3	0.9175	0.9376	0.9319	0.9309	0.9295
Unet++	0.9279	0.9348	**0.9585**	0.9191	**0.9351**
DeepLabV3+	**0.9283**	0.9348	0.9572	0.9183	0.9347
FPN	0.9220	**0.9399**	0.9383	**0.9319**	0.9330

Table 4. Methods Evaluations for Specificity Metric

Methods	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Average**
Unet	0.9632	0.9566	**0.9641**	0.9538	**0.9594**
DeepLabV3	0.9632	0.9511	0.9615	0.9474	0.9558
Unet++	0.9599	0.9579	0.9571	0.9548	0.9574
DeepLabV3+	0.9607	**0.9595**	0.9557	**0.9550**	0.9577
FPN	**0.9633**	0.9519	0.9637	0.9481	0.9567

Table 5. Methods Evaluations for Accuracy Metric

Methods	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Average**
Unet	**0.9457**	0.9530	**0.9573**	0.9517	**0.9519**
DeepLabV3	**0.9457**	0.9531	0.9547	0.9501	0.9509
Unet++	0.9423	0.9527	0.9562	**0.9536**	0.9512
DeepLabV3+	0.9436	**0.9581**	0.9542	0.9533	0.9515
FPN	0.9430	0.9536	0.9564	0.9503	0.9508

As can be seen from the tables, the segmentation performance of the utilized deep learning architectures was evaluated across four distinct scenarios, each presenting unique training and testing conditions. The effectiveness of each model was assessed based on five primary metrics: IoU, Dice coefficient, sensitivity, specificity, and accuracy. The results demonstrate that model performance varies significantly across different dataset compositions, indicating the impact of data diversity and source-specific characteristics on segmentation quality.

In Scenario 1, where models were trained on the public dataset and tested on the private dataset, the DeepLabV3 model achieved the highest IoU score of 0.7524. This score suggests that DeepLabV3 can capture lesion boundaries effectively, even when applied to data that differ in distribution from the training set. In Scenario 2, which involved training and testing on a hybrid dataset combining public and private data, FPN outperformed other models, achieving an IoU score of 0.8146. This result highlights the benefit of mixed data training, as FPN could adapt to both public and private image features, improving generalization. Scenario 3, trained and tested solely on private data, Unet yield the best IoU performance, with a score of 0.8075. In Scenario 4, training and testing were conducted on the public dataset. Unet++ achieved the top IoU of 0.8224, likely due to the model's strong alignment with the diversity found within public datasets.

For Dice coefficients, the DeepLabV3 model attained the highest result of 0.8636 in Scenario 1, indicating superior overlap between predicted and ground truth lesion areas on private data. This shows DeepLabV3's potential to generalize well, even when the test set differs from the training set. In Scenario 2, FPN exhibited the highest Dice coefficient of 0.8939, benefiting from the hybrid dataset's balanced diversity and enhancing the segmentation accuracy. In Scenario 3, with only private data, FPN again performed best, achieving a Dice score of 0.8874, affirming its effectiveness in segmenting private data lesions. In Scenario 4, where the public dataset was used exclusively, Unet++ achieved a Dice score of 0.8859, suggesting a robust performance on homogeneous data sources.

Across all scenarios, Unet++ showed the highest sensitivity in Scenario 3 (0.9581), indicating its strong ability to identify true positive lesion pixels on private data. This suggests that Unet++ is particularly adept at capturing lesion regions within private data, which often include complex lesion shapes. Conversely, in Scenario 2, DeepLabV3 and FPN also demonstrated high sensitivity values, suggesting that these models can successfully detect true lesion areas across varying data sources.

The Unet model recorded the highest specificity in Scenario 3, reaching 0.9641. This metric reflects Unet's ability to minimize false positives in segmenting lesions from private dataset backgrounds, thus distinguishing lesion boundaries with greater precision. In Scenario 1, DeepLabV3 and FPN models also achieved commendable specificity scores, highlighting their capacity to differentiate lesion regions accurately, even with only public dataset training.

Regarding accuracy, DeepLabV3+ provided the most consistent performance across all scenarios, with its highest accuracy recorded in Scenario 3 (0.9564). This indicates that DeepLabV3+ effectively balances lesion detection and segmentation accuracy, especially in private datasets. The hybrid dataset in Scenario 2 also led to solid accuracy scores across models, reflecting the advantages of mixed data training for segmentation tasks.

Overall, segmentation models exhibited the robust performance across multiple metrics, indicating their potential for clinical applications requiring reliable skin lesion segmentation. These findings emphasize the importance of diverse training datasets and suggest that models trained on hybrid data are more adaptable to real-world dermatological applications where public and private data characteristics may be present.

One of the key advantages of the proposed study is the inclusion of a unique database comprising images obtained from patients at the Dermatology Clinic of Sakarya University Education and Research Hospital under ethically approved protocols. This database combines open-source and private data, enabling a robust analysis across diverse data sources. Our study addresses these limitations by incorporating diverse lesion types and varied imaging conditions from private clinical data, enabling a more robust and comprehensive model evaluation. Working with a hybrid database enhances the generalizability of segmentation models and strengthens their performance in real-time applications. Additionally, the study offers a comparative analysis of segmentation performance across widely recognized architectures, such as Unet, DeepLabV3, Unet++, DeepLabV3+, and FPN. This approach allows for identifying the highest-performing architecture, providing a benchmark for evaluating segmentation accuracy compared to existing studies.

Finally, Table 6 compares the results obtained with those of existing studies. The study in [32] proposed a melanoma segmentation approach, including Unet and LinkNet deep learning networks, coupled with transfer learning and fine-tuning techniques. Experiments conducted on three publicly available databases (PH2, ISIC 2018, and DermIS) have shown promising results, with Unet demonstrating notable performance. Specifically, the average Dice coefficient achieved was 0.923 on the PH2 database, 0.893 on ISIC 2018, and 0.879 on the DermIS database. These findings indicate significant success for U-net across the evaluated databases. The study in [11] introduced a hyper-parameter optimized Fully Convolutional Encoder-Decoder Network (FCEDN) for dermoscopy image segmentation. The novel Exponential Neighborhood Grey Wolf Optimization (EN-GWO) algorithm was employed to optimize network hyper-parameters. EN-

GWO incorporates a neighborhood-based searching strategy, combining individual and global haunting strategies of wolves to strike a balance between exploration and exploitation. The study compared EN-GWO with four variants of Grey Wolf Optimization (GWO), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) for hyperparameter optimization on the ISIC 2016 and 2017 databases. The proposed model achieves a Jaccard index of 96.41% and 86.85%, a Dice coefficient of 98.48% and 87.23%, and an accuracy of 98.32% and 95.25% for the ISIC 2016 and 2017 databases.

Table 6. Comparison of the Proposed System with Recent Studies

Reference No	Data	Model	Dice	IoU
[32]	ISIC 2018 PH2 DermIS	U-net + LinkNet	0,8940	0,8090
[33]	ISIC 2016 ISIC 2017	FCEDN	0,9848 0,8723	0,9641 0,8685
Ours Best	Hybrid	FPN	0,8939	0,8146

The proposed segmentation approach also gives promising results compared to the other studies. Unlike studies that rely solely on public datasets, the study’s dataset includes 3463 public images and 765 private clinical images, enhancing the robustness and generalizability of the given model. Moreover, the comprehensive evaluation strategy of the study, involving four distinct training/test combinations, ensures the results' reliability and robustness. This thorough approach contrasts with the more limited evaluation methodologies reported in the earlier studies.

The results obtained in this study outperform those of earlier works, particularly in the hybrid dataset scenario. Studies such as [33] relied on GAN-based architectures but struggled with small-scale lesions. Our use of FPN and DeepLabV3+ provides better handling of multi-scale lesions, as demonstrated by the superior IoU and Dice scores. However, one limitation of our approach is the reliance on a relatively small private dataset, which may limit the model’s generalizability. Future work should focus on expanding the private dataset and exploring transformer-based models for improved segmentation accuracy.

5. Conclusions

This study comprehensively evaluated deep learning architectures for skin lesion segmentation using a unique hybrid dataset comprising public ISIC images and private clinical data. By employing multiple segmentation models, including Unet, Unet++, DeepLabV3, DeepLabV3+, and FPN, across different training and testing scenarios, this work highlighted the impact of data diversity on segmentation performance. The findings indicate that integrating public and private datasets during training significantly enhances model generalizability and segmentation accuracy, especially in heterogeneous clinical applications.

The results demonstrate that the FPN and DeepLabV3+ models consistently perform well across critical metrics such as IoU, Dice coefficient, sensitivity, specificity, and accuracy. FPN, in particular, excelled in hybrid dataset scenarios, suggesting that it effectively adapts to varying lesion types and imaging conditions, an essential characteristic for practical applications in dermatology. These models' ability to delineate lesion boundaries accurately, even within challenging image contexts, underscores their potential suitability for deployment in automated skin lesion analysis tools.

Integrating private clinical data strengthens model robustness and addresses limitations inherent in publicly available datasets, which often need more complexity and specificity of clinical images. This study underscores the importance of hybrid datasets for training deep learning models in medical imaging, as they contribute to more reliable and clinically applicable segmentation outcomes.

Future work should focus on expanding the dataset to include a wider variety of lesion types and further investigate transformer-based or hybrid architectures that can enhance long-range dependency modeling in segmentation. Additionally, attention-based mechanisms could be explored to refine boundary detection in challenging cases. The promising results of this study support the advancement of automated diagnostic tools in dermatology, facilitating early detection and improved treatment planning for skin cancer patients.

References

[1] S. Spanos *et al.*, “Measuring the quality of skin cancer management in primary care: A scoping review,” *Australas. J. Dermatol.*, vol. 64, no. 2, pp. 177–193, May 2023, doi: 10.1111/AJD.14023.

[2] R. Javed, M. S. M. Rahim, T. Saba, and A. Rehman, “A comparative study of features selection for skin lesion detection from dermoscopic images,” *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 9, no. 1, pp. 1–13, Dec. 2020, doi: 10.1007/S13721-019-0209-1/TABLES/5.

[3] M. Zafar, M. I. Sharif, M. I. Sharif, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, “Skin Lesion Analysis and Cancer Detection Based on Machine/Deep Learning Techniques: A Comprehensive Survey,” *Life 2023, Vol. 13, Page 146*, vol. 13, no. 1, p. 146, Jan. 2023, doi: 10.3390/LIFE13010146.

[4] Z. Mirikharaji *et al.*, “A survey on deep learning for skin lesion segmentation,” *Med. Image Anal.*, vol. 88, p. 102863, Aug. 2023, doi: 10.1016/J.MEDIA.2023.102863.

[5] N. C. F. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2018-April, pp. 168–172, Oct. 2017, doi: 10.1109/ISBI.2018.8363547.

[6] N. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," Feb. 2019, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/1902.03368v2.

[7] C. Hernández-Pérez *et al.*, "BCN20000: Dermoscopic Lesions in the Wild," *Sci. Data*, vol. 11, no. 1, Aug. 2019, doi: 10.1038/s41597-024-03387-w.

[8] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data 2018 51*, vol. 5, no. 1, pp. 1–9, Aug. 2018, doi: 10.1038/sdata.2018.161.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[10] N. J. Singh and K. Nongmeikapam, "Semantic Segmentation of Satellite Images Using Deep-Unet," *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 1193–1205, Feb. 2023, doi: 10.1007/S13369-022-06734-4/TABLES/2.

[11] L. Zhang, J. Shen, and B. Zhu, "A research on an improved Unet-based concrete crack detection algorithm," *Struct. Heal. Monit.*, vol. 20, no. 4, pp. 1864–1879, Jul. 2021, doi: 10.1177/1475921720940068/ASSET/IMAGES/10.1177_1475921720940068-IMG1.PNG.

[12] D. Harrison, F. C. De Leo, W. J. Gallin, F. Mir, S. Marini, and S. P. Leys, "Machine Learning Applications of Convolutional Neural Networks and Unet Architecture to Predict and Classify Demosponge Behavior," *Water 2021, Vol. 13, Page 2512*, vol. 13, no. 18, p. 2512, Sep. 2021, doi: 10.3390/W13182512.

[13] D.-Y. Chen *et al.*, "Building Extraction and Number Statistics in WUI Areas Based on UNet Structure and Ensemble Learning," *Remote Sens. 2021, Vol. 13, Page 1172*, vol. 13, no. 6, p. 1172, Mar. 2021, doi: 10.3390/RS13061172.

[14] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, 2018, doi: 10.1007/978-3-030-00889-5_1/FIGURES/3.

[15] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 833–851, Feb. 2018, doi: 10.1007/978-3-030-01234-2_49.

[16] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," Apr. 2018, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/1804.03999v3.

[17] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/J.MEDIA.2019.101552.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Nov. 2015, doi: 10.1109/TPAMI.2016.2644615.

[19] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Nov. 2014, doi: 10.1109/TPAMI.2016.2572683.

[20] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," Feb. 2018, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/1802.06955v5.

[21] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 565–571, Jun. 2016, doi: 10.1109/3DV.2016.79.

[22] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," *Neural Networks*, vol. 121, pp. 74–87, Feb. 2019, doi: 10.1016/j.neunet.2019.08.025.

[23] H. Sharen, M. Jawahar, L. Jani Anbarasi, V. Ravi, N. Saleh Alghamdi, and W. Suliman, "FDUM-Net: An enhanced FPN and U-Net architecture for skin lesion segmentation," *Biomed. Signal Process. Control*, vol. 91, p. 106037, May 2024, doi: 10.1016/J.BSPC.2024.106037.

[24] Sweta Jain, Pruthviraj Choudhari, Mahesh Gour, Pulmonary Lung Nodule Detection from Computed Tomography Images Using Two-Stage Convolutional Neural Network, *The Computer Journal*, Volume 66, Issue 4, April 2023, Pages 785–795.

[25] He, X., Wang, Y., Poiesi, F., Song, W., Xu, Q., Feng, Z., & Wan, Y. (2023). Exploiting multi-granularity visual features for retinal layer segmentation in human eyes. *Frontiers in Bioengineering and Biotechnology*, *11*, 1191803.

[26] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

[27] Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis, 58,* 101552. https://doi.org/10.1016/j.media.2019.101552

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M.,

Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[29] M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, “A survey, review, and future trends of skin lesion segmentation and classification,” *Comput. Biol. Med.*, vol. 155, p. 106624, Mar. 2023, doi: 10.1016/J.COMPBIOMED.2023.106624.

[30] M. Strzelecki, M. Kociołek, M. Strąkowska, M. Kozłowski, A. Grzybowski, and P. M. Szczypiński, “Artificial intelligence in the detection of skin cancer: State of the art,” *Clin. Dermatol.*, vol. 42, no. 3, pp. 280–295, May 2024, doi: 10.1016/J.CLINDERMATOL.2023.12.022.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” Dec. 2016, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/1612.03144v2.

[32] R. L. Araújo, F. H. D. d. Araújo, and R. R. V. e. Silva, “Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning,” *Multimed. Syst.*, vol. 28, no. 4, pp. 1239–1250, Aug. 2022, doi: 10.1007/S00530-021-00840-3/TABLES/8.

[33] R. Mohakud and R. Dash, “Skin cancer image segmentation utilizing a novel EN-GWO based hyper-parameter optimized FCEDN,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9889–9904, Nov. 2022, doi: 10.1016/J.JKSUCI.2021.12.018.

Article Information Form

Authors Contributions

Dr. Gökçen ÇETİNEL: Conceptualization, study design, critical revision of the manuscript, final editing, supervision, and project administration.
Bekir Murat Aydın: Software development and data collection
Dr. Sevda GÜL: Literature review, software development, results interpretation, and data validation
Dr. Devrim AKGÜN: Methodology development, software development, and study design.
Dr. Rabia ÖZTAŞ KARA: Data collection, data validation and statistical analysis.

Conflict of Interest Notice

Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study is supported by The Scientific and Technical Research Council of Turkey (TUBITAK) through The Research Support Programs Directorate (ARDEB) with project number of 122E629.

Ethical Approval

The Ethical Committee of Sakarya University Medical Sciences Faculty approved the study. All procedures in studies involving human participants were by the ethical standards of the institutional and/or national search committee.

Availability of Data and Materials

No Applicable

Plagiarism Statement

This article has been scanned by iThenticate^{TM}.