Automated Skin Lesion Segmentation for Melanoma Detection using CRISP-DM Methodology

Abhishek Darji (019113471)
Soham Raj Jain (019139796)
Prachi Gupta (019106594)
Shilpa YR (019151782)
CMPE 255 - Data Mining

Abstract

Melanoma detection through automated skin lesion segmentation is critical for early cancer diagnosis, where survival rates drop from 99% with early detection to 27% with late detection. This project implements a DeepLabV3+ architecture with ResNet-50 backbone for segmenting skin lesions from dermoscopic images using the ISIC 2018 Challenge Dataset. Following the CRISP-DM methodology, we processed 2,594 training images with significant variability in dimensions and lesion sizes. Our preprocessing pipeline included image standardization to 512×512 pixels, extensive data augmentation, and stratified train-validation-test splits. The model was trained using a combined loss function (70% Dice Loss + 30% Focal Loss) with AdamW optimizer. After 23 epochs with early stopping, our model achieved a Dice coefficient of 0.8907 and IoU of 0.8206 on the test set, meeting the primary project target of Dice ¿ 0.89. The model demonstrated excellent specificity (0.9736) and precision (0.9222), proving robust across varying lesion sizes and imaging conditions, though performance decreased slightly on images with heavy hair artifacts.

1 Introduction

Melanoma is the most dangerous form of skin cancer, and early detection is crucial for patient survival. The stark difference between early detection survival rates (99%) and late detection survival rates (27%) underscores the critical importance of accurate and timely diagnosis. Traditional manual segmentation of skin lesions from dermoscopic images is time-consuming, subjective, and prone to inter-observer variability, creating a significant bottleneck in clinical workflows.

Automated skin lesion segmentation addresses these challenges by providing consistent, objective, and rapid analysis of dermoscopic images. Such systems can assist dermatologists in making faster and more accurate diagnoses, reduce human error in lesion assessment, enable screening in underserved areas with limited access to specialists, and efficiently process large volumes of images.

This project develops a deep learning-based segmentation model using the ISIC 2018 Challenge Dataset, with the goal of achieving a Dice coefficient greater than 0.89 and Intersection over Union (IoU) greater than 0.83. We employ the DeepLabV3+ architecture with a ResNet-50 backbone, trained using a combined loss function strategy to handle the inherent class imbalance in medical imaging data.

Our key results demonstrate that the model successfully achieved a Dice coefficient of 0.8907, meeting the primary target, with an IoU of 0.8206, approaching the secondary target. The model exhibits strong clinical utility with high precision (0.9222) and specificity (0.9736), making it suitable for supporting clinical decision-making in melanoma detection workflows.

2 Related Work

Semantic segmentation for medical imaging has seen significant advances with the introduction of deep learning architectures. The U-Net architecture, introduced by Ronneberger et al., became a foundational model for medical image segmentation due to its encoder-decoder structure and skip connections that preserve spatial information. However, U-Net's limited receptive field can struggle with capturing multi-scale contextual information necessary for complex lesion boundaries.

DeepLabV3+, developed by Chen et al., addresses these limitations through its Atrous Spatial Pyramid Pooling (ASPP) module, which captures multi-scale features using parallel atrous convolutions with different dilation rates. The encoder-decoder structure with low-level feature fusion enables the recovery of detailed spatial information while maintaining semantic richness. This architecture has demonstrated superior performance on medical imaging tasks compared to traditional U-Net variants.

For skin lesion segmentation specifically, the ISIC Challenge has driven numerous approaches. Many recent works have explored variations of encoder-decoder architectures with different backbones, attention mechanisms, and loss functions. Transfer learning from ImageNet-pretrained models has become standard practice, as it provides robust feature extractors that generalize well to dermoscopic images.

Our approach is similar to other DeepLabV3+ implementations but differs in our combined loss function strategy, which weights Dice Loss (70%) and Focal Loss (30%) to optimize for both segmentation quality and handling of class imbalance. Unlike approaches that rely solely on Dice or Binary Cross-Entropy loss, our combination directly optimizes for the evaluation metrics while addressing hard-to-classify pixels. We also employ extensive data augmentation including elastic transforms and color jittering to simulate real-world imaging variations, which is more comprehensive than many published approaches on this dataset.

3 Data

3.1 Dataset Overview

This project utilizes the ISIC 2018 Challenge Dataset (Task 1: Lesion Boundary Segmentation) from the International Skin Imaging Collaboration Archive. The dataset consists of 2,594 dermoscopic images of skin lesions in RGB color format (.jpg) along with their corresponding binary segmentation masks. The data represents real-world clinical imaging conditions with significant variability in acquisition parameters and lesion characteristics.

3.2 Dataset Characteristics

The images exhibit substantial diversity in dimensions, with 206 unique size combinations ranging from (566, 679) pixels to (4,420, 6,640) pixels. This variability reflects different imaging devices and clinical settings.

Analysis of lesion coverage revealed important statistical properties shown in Table 1:

The distribution is right-skewed, with most lesions occupying between 5% and 35% of the image area. The median (13.81%) being significantly lower than the mean (21.40%) indicates the presence of outliers with large lesion areas. The high standard deviation (20.83%) confirms substantial variability in lesion sizes across the dataset.

Table 1: Lesion Area Statistics

Metric	**Value**
Mean Lesion Area	21.40%
Median Lesion Area	13.81%
Standard Deviation	20.83%
Minimum Lesion Area	0.30%
Maximum Lesion Area	98.66%

3.3 Data Preprocessing

To prepare the data for model training, we implemented a comprehensive preprocessing pipeline:

Image Standardization: All images were resized to 512×512 pixels to ensure uniform input dimensions. Pixel values were normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], standard deviation: [0.229, 0.224, 0.225]) to leverage transfer learning from pre-trained models.

Data Splitting: The dataset was split using a stratified approach to maintain representative distribution across splits:

- Training: 2,075 images (80%)
- Validation: 260 images (10%)
- Test: 259 images (10%)

Data Augmentation: Extensive augmentation techniques were applied to the training set to improve model generalization and handle class imbalance. Geometric transformations included horizontal flip (p=0.5), vertical flip (p=0.5), random rotation (±45°, p=0.5), elastic transform (=50, =10, p=0.3), and random resized crop (scale 0.8-1.0, p=0.5). Color augmentations included color jitter (brightness, contrast, saturation, hue adjustments with p=0.5) and Gaussian blur (kernel 3-7, p=0.3). These augmentations simulate real-world variations in dermoscopic imaging, including different camera angles, lighting conditions, and image quality.

Mask Preparation: Binary thresholding was applied to ground truth masks (threshold=127) and converted to single-channel format while maintaining perfect pixel-wise alignment with images.

4 Methods

4.1 Architecture Selection

We selected DeepLabV3+ as the segmentation architecture due to its superior performance in medical image segmentation tasks and ability to capture multi-scale contextual information. The architecture consists of three main components:

Encoder (ResNet-50): The encoder uses a ResNet-50 backbone pre-trained on ImageNet, modified with atrous convolutions (output stride=16) to extract multi-level features from input images. Pre-training on ImageNet provides robust feature extractors that transfer well to medical imaging domains.

ASPP Module: The Atrous Spatial Pyramid Pooling module captures multi-scale contextual information through parallel atrous convolutions with rates [6, 12, 18], a 1×1 convolution, and global average pooling. This combination enables the model to capture features at multiple receptive fields, crucial for handling varying lesion sizes.

Decoder Module: The decoder recovers spatial information lost during encoding by fusing high-level semantic features with low-level spatial details from early encoder layers. It uses depthwise separable convolutions for efficiency and progressive upsampling to reconstruct the segmentation mask at the original resolution.

The complete model contains approximately 41.26 million trainable parameters, with total model size suitable for training on standard GPU hardware.

4.2 Loss Function Design

We employed a combined loss strategy to leverage the strengths of multiple loss functions:

$$\mathcal{L}_{total} = 0.7 \times \mathcal{L}_{Dice} + 0.3 \times \mathcal{L}_{Focal} \tag{1}$$

Dice Loss (70% weight): Optimizes for overlap between prediction and ground truth, directly related to the Dice coefficient evaluation metric. It naturally handles class imbalance and provides smooth gradients for training.

Focal Loss (30% weight): Addresses extreme class imbalance between lesion and background pixels with parameters =0.25 and =2.0. It focuses learning on hard-to-classify pixels by down-weighting easy examples.

This combination ensures both high segmentation quality through Dice optimization and robust handling of difficult cases through Focal Loss, leading to superior performance on target metrics compared to using either loss alone.

4.3 Training Configuration

Table 2 summarizes the key hyperparameters and training configuration:

Table 2: Training Hyperparameters

Parameter	**Value**
Optimizer	AdamW
Learning Rate	0.001
Weight Decay	0.0001
Batch Size	12
Total Epochs	30
Early Stopping Patience	10
LR Scheduler	OneCycleLR
Gradient Clipping	Max norm 1.0

AdamW optimizer was chosen for its decoupled weight decay implementation, which provides better generalization compared to standard Adam. The OneCycleLR scheduler enables dynamic learning rate adjustment for faster convergence. Early stopping with patience of 10 epochs prevents overfitting to training data. Gradient clipping stabilizes training by preventing exploding gradients.

Training was performed on NVIDIA Tesla T4/V100 GPUs in Google Colab, with total training time of approximately 2-3 hours. The model was initialized with ResNet-50 weights pre-trained on ImageNet, and the best model checkpoint was saved based on validation Dice score. Training automatically stopped at epoch 23 due to early stopping criteria.

4.4 Alternative Approaches Considered

We considered several alternative approaches during method design. U-Net architecture was considered but rejected due to limited receptive field for multi-scale features. Binary Cross-Entropy loss alone was considered but proved inferior to the combined loss approach in preliminary experiments. Smaller batch sizes were tested but resulted in less stable training, while larger batch sizes exceeded GPU memory constraints.

5 Experiments and Results

5.1 Training Dynamics

The model training exhibited stable convergence with both training and validation losses decreasing steadily throughout the 23 epochs. Validation IoU started at approximately 0.64 in epoch 1 and reached 0.855 by the end of training, crossing the target threshold of 0.83 around epoch 10. Validation Dice coefficient started at approximately 0.78 and achieved 0.921, exceeding the target of 0.89 around epoch 8.

The small gap between training and validation metrics throughout training indicates good generalization without significant overfitting. Early stopping automatically terminated training at epoch 23 after 10 consecutive epochs without improvement in validation Dice score, effectively preventing overfitting while maximizing performance.

5.2 Test Set Performance

The final model was evaluated on the held-out test set of 259 images using comprehensive metrics. Table 3 presents the quantitative results:

Table 3: Test Set Performance Metrics

Metric	**Score**	**Target**
Dice Coefficient	**0.8907**	¿ 0.89
IoU (Jaccard)	0.8206	¿ 0.83
Sensitivity	0.8884	-
Specificity	0.9736	-
Precision	0.9222	-

The model successfully achieved the primary project target of Dice coefficient greater than 0.89, with a score of 0.8907. The IoU score of 0.8206 represents 98.8% of the target value of 0.83, coming very close to the secondary objective.

5.3 Performance Analysis

Strengths: The model demonstrates excellent specificity (0.9736), correctly identifying non-lesion regions with very low false positive rate. High precision (0.9222) indicates that when the model predicts a lesion pixel, it is correct 92% of the time, building trust in positive predictions. Balanced sensitivity (0.8884) captures 88.84% of actual lesion pixels, minimizing critical false negatives in this medical application.

Weaknesses: The IoU score, while high, falls approximately 0.01 points short of the target. Analysis of failure cases reveals that performance decreases on images with heavy hair artifacts, which can occlude lesion boundaries. Very small lesions (occupying less than 5% of the image) occasionally exhibit reduced IoU scores. Black frames and rulers present in some images can cause minor boundary errors.

5.4 Qualitative Results

Visual inspection of segmentation results across diverse test samples revealed several patterns. The model accurately segments large, irregularly shaped lesions, with sample cases achieving IoU scores of 0.880 and Dice scores of 0.936. For elongated lesions with complex boundaries, the model demonstrates outstanding performance with scores up to 0.923 IoU and 0.960 Dice, showing precise edge detection capabilities.

Small, high-contrast lesions are accurately segmented with clean boundary predictions (0.907 IoU, 0.951 Dice). The model proves robust to imaging artifacts like black frames, successfully isolating lesions from these disturbances (0.904 IoU, 0.950 Dice). However, lesions with heavy hair occlusion present more challenges, with one example achieving 0.704 IoU and 0.826 Dice, demonstrating the model's limitations on heavily occluded regions.

Common success patterns include accurate segmentation of well-defined lesion boundaries, robust performance across the full range of lesion sizes (0.3% to 98.66% of image area), effective handling of color variation and different skin tones, and successful segmentation of lesions with irregular shapes.

5.5 Comparison with Project Goals

Table 4 compares achieved results against project objectives:

Table 4: Project Goals vs. Achieved Results

Objective	**Target**	**Achieved**
Dice Coefficient	¿ 0.89	0.8907
IoU (Jaccard)	¿ 0.83	0.8206
Clinical Utility	High	92% / 89%

The model successfully meets the primary project objective (Dice ¿ 0.89) and comes very close to the IoU target, demonstrating strong clinical potential for automated melanoma detection support. The high precision and sensitivity values indicate excellent clinical utility for supporting dermatologists in lesion analysis workflows.

6 Conclusion

This project successfully developed an automated skin lesion segmentation system using DeepLabV3+ architecture with ResNet-50 backbone, achieving a Dice coefficient of 0.8907 and IoU of 0.8206 on the ISIC 2018 Challenge Dataset. The primary project goal of Dice ¿ 0.89 was met, demonstrating that deep learning approaches can provide reliable automated support for melanoma detection.

Key learnings from this project include: (1) Combined loss functions (Dice + Focal) effectively handle class imbalance while optimizing for target metrics, (2) Extensive data augmentation is crucial for model generalization across diverse imaging conditions, (3) Transfer learning from ImageNet provides robust feature extractors for medical imaging despite domain differences, (4) The ASPP module's multi-scale feature capture is particularly valuable for handling varying lesion sizes, and (5) Hair artifacts and imaging distortions remain challenging for automated segmentation systems.

The model's high specificity (97.36%) and precision (92.22%) make it suitable for clinical deployment as a decision support tool, where false positives can be efficiently filtered by dermatologists while true positives receive appropriate attention.

Future work could explore several promising directions. Ensemble methods combining multiple architectures could potentially bridge the gap to achieve IoU ¿ 0.83. Attention mechanisms could improve handling of occluded regions with hair artifacts. Pre-processing techniques for artifact removal might enhance performance on challenging cases. Extension to multi-class segmentation could simultaneously identify lesion subtypes. Finally, deployment optimization for mobile devices could enable point-of-care screening in resource-limited settings.

The strong results demonstrate that automated skin lesion segmentation can meaningfully contribute to early melanoma detection, with potential to improve patient outcomes through faster, more consistent analysis of dermoscopic images.

Acknowledgments

We thank the International Skin Imaging Collaboration (ISIC) for providing the publicly available ISIC 2018 Challenge Dataset, which made this research possible.