

Diagnosis-Stratified U-Net for Dermoscopic Lesion Segmentation: A Reproducible Baseline with Boundary-Aware Evaluation

Ms. B. Bhavya Sri^{1*}, Dr. Vijay Kumar Burugari²

¹Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
Mobile: +91 9985452118

Email: bhimabhavyasri@gmail.com

²Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

Mobile: +91 7032599767

Email: vijaybru@gmail.com

*Corresponding Author: Ms. B. Bhavya Sri, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

Email: bhimabhavyasri@gmail.com

ABSTRACT

Background:

Automated skin lesion segmentation is essential to computer-aided diagnosis, with existing approaches commonly neglecting boundary accuracy and performance variability based on morphology, which constrains clinical applications to surgical planning.

Methods:

We introduce a diagnosis-conscious U-Net model tested on the ISIC 2018 dataset (519 test images) with stratified partitioning to avoid diagnostic leakage. The pipeline uses error-graded augmentation and end-to-end assessment based on overlap measures (Dice, IoU), boundary-sensitive Hausdorff distance, and per-diagnosis assessment between melanoma, nevus, and seborrheic keratosis. A systematic comparison is made between baseline configurations and stratified splitting, full augmentation and Dice loss optimization in an ablation study.

Results:

The model attained a Dice coefficient of 0.871 ± 0.138 , sensitivity of 0.921 and specificity of 0.968. The mean Hausdorff distance was 24.65 ± 28.94 pixels (~ 1.23 mm) which showed varying boundary accuracy. The heterogeneity of performance was demonstrated with per-diagnosis analysis: nevus (0.890), melanoma (0.840), and seborrheic keratosis (0.779 ± 0.230). The stratified splitting enhanced the minority-class F1 by 0.010, whereas aggressive augmentation harmed the performance, implying that the optimal performance requires dataset-specific optimization.

Conclusion:

In morphology-sensitive failure modes that are masked by aggregate measures, diagnosis-stratified evaluation can be used. Although it is appropriate to screen, the 1.23 mm deviation in a boundary requires adjustment to margin-critical surgical uses. To facilitate reproducible benchmarking and improvements in the future with morphology-based adjustments, we publish code and pretrained weights.

Keywords: Skin Lesion Segmentation, Dermoscopic Imaging, U-Net Architecture, Diagnosis-Stratified Splitting, Boundary-Aware Evaluation, Reproducible Medical AI

How to cite this article: B. Bhavya Sri^{1*} Dr. Vijay Kumar Burugari². Diagnosis-Stratified U-Net for Dermoscopic Lesion Segmentation: A Reproducible Baseline with Boundary-Aware Evaluation. *Int J Drug Deliv Technol.* 2026;16(86s):1458-1469. DOI: 10.25258/ijddt.16.86s.126.

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

Skin cancer is one of the most widespread causes of dermatologic death among the global population,

and the timely detection of lesions is significantly based on the correct delineation of the lesion on the basis of dermoscopic images [4]. Computer-aided

diagnosis systems rely on automated segmentation as the preprocessing stage of the subsequent classification, morphological analysis, and surgical planning [5]. Although encoder-decoder structures, especially U-Net variants, have solid baselines in biomedical segmentation, their clinical implementation is limited by the lack of a consistent evaluation procedure and consideration of boundary precision, which is directly proportional to excision margin accuracy [21].

The existing body of literature mostly maximizes the metrics of region-based overlap like Dice and IoU that tends to hide localization errors of important boundaries. Nordstrom et al. showed that loss function, and binarization threshold have a major effect on segmentation fidelity but most benchmarks still report aggregate metrics without boundary-aware validation [17]. In addition, random split of datasets in medical images often creates patient, or diagnosis, level leakage, which artificially inflates reported performance. Hoang et al. introduced the use of class imbalance as categorical weighting schemes, which does not consider morphology-specific generalization between lesion subtypes [6]. New adaptive augmentation strategies have been found to enhance morphological variability robustness. To retain diagnostic information whilst improving diversity, Kim et al. suggested cyclic intensity modulation [12], and Verma et al. combined confidence-guided curriculum learning [22]. Nevertheless, these approaches typically need multi-model ensembles or hand difficulty labeling, which adds calculation burden that can make them deployment infeasible. Equally, the architectural advances like dual-encoder U-Nets [11] and attention-enhanced bottlenecks have shown better overlap metrics but seldom give per-diagnosis performance or edge deviation, and in any case, morphology-dependent failure modes are not characterized.

In response to these drawbacks, we introduce a diagnosis-sensitive segmentation architecture that emphasizes reproducible benchmarking and boundary accuracy and specific regularization instead of complexity in architecture. Our algorithm presents three design ideas: (i) diagnosis-stratified partitioning of the dataset to apply distributional alignment and avoid leakage, (ii) an error-guided hard-example augmentation operator that dynamically focuses the gradient updates on boundary-ambiguous samples, and (iii) a thorough evaluation protocol reporting per-diagnosis Dice and the 95th-percentile Hausdorff distance to measure contour deviation. In contrast to ensemble-based variants [5], our method uses a streamlined U-Net backbone that is trained on differentiable Dice loss [7], making it computationally efficient without confounding the effects of the data-level regularization.

The main contributions of this work are:

1. A diagnosis-stratified splitting scheme that alleviates class-conditional distribution shift when validating and testing.
2. Lightweight, error-guided augmentation strategy that uses implicit curriculum learning in the absence of multi-model inference.
3. A boundary-conscious assessment system reporting per-diagnosis DSC, IoU and Hausdorff distance with careful consideration of morphology dependent variance.
4. Publication of the complete code and pretrained weights and configuration manifests such that they can be replicated independently.

II. RELATED WORK:

The art of automated skin lesion segmentation in dermoscopy has gone through changes, and encoder-decoder networks have become the paradigm of the pixel-wise lesion delineation. Initial standards showed that the traditional convolutional networks were competitive in overlap measures on publicly available datasets [4]. Nevertheless, the quest to obtain incremental performance improvements has excessively biased the literature in favor of architecturally rich variants, frequently to the detriment of reproducibility, clinical interpretability and stringent evaluation protocols. We contextualize our methodology in four dimensions that are critical to it, namely, architectural design, loss optimization under imbalance, adaptive augmentation strategies, and boundary-aware evaluation.

Complexity Trade-offs and Architectural Baselines. Canonical U-Net architecture has been the fact of the biomedical segmentation since it has symmetric skip connections and effective feature propagation. This topology has been heavily altered by recent works to include multi-scale dependencies. An example is attention-augmented bottlenecks [1], which directly combine transformer-based global context with local convolutional features, and marginal Dice improvements on ISIC benchmarks. Combination frameworks of Mask R-CNN and DeepLabV3+ [5] continue to drive the overlap scores even higher at the cost of significant computational load and deployment complexity. These multi-encoder or attention-intensive designs are flexible in architecture, but tend to obscure the marginal utility of every constituent and do not have a normalized report on boundary precision or more diagnostic generalisation. Conversely, our pipeline has a slender U-Net backbone, which focuses on architectural transparency and reproducible training dynamics, rather than structural complexity.

Loss Functions and Class Imbalance Mitigation. Medical segmentation also has a critical foreground-background skew, making pixel-wise cross-entropy not optimal to lesion delineation. The soft-Dice loss has emerged as the most common alternative as it directly optimizes spatial overlap [17]. But standard Dice formulations do not make the distinction between the foreground and background pixels, and

so ignore the imbalance in difficulty where the gradient update is concentrated in a boundary-ambiguous or morphologically irregular region. Recent work goes here through pixel-wise modulation [7] or categorical weighting schemes that are more rough on minority-class misclassifications [6]. Our results incorporate the differentiable Dice loss as the overall optimization target, which is consistent with the existing medical imaging workflow, but explicitly pairs this with diagnosis-stratified data sampling to guarantee the proportional representation of rare lesion types (i.e., melanoma) when constructing a batch, avoiding distributional bias without any auxiliary loss terms. Adaptive Augmentation and Curriculum Strategies. The baseline regularization offered by static augmentation pipelines (random flips, rotations, color jitter) does not dynamically attack the model weaknesses. The learning of curriculum and hard-example mining have been proposed in order to gradually subject networks to high-entropy samples. Other methods like confidence-guided multi-model refinement [22] and gradually recurrent supervision [24] separate the training data in terms of lesion complexity, so that easy samples do not dominate the early optimization epochs. Generalization on structurally varied medical images has also been demonstrated to be better on cyclic schedules of perturbation augmentation that dynamically vary the perturbation intensity [12]. Our pipeline, based on these principles, uses a lightweight error-guided augmentation operator, which applies intensive deformations (elastic transforms, coarse dropout) when per-sample Dice scores are below a predetermined threshold. This dynamic gating focuses gradient updates on ambiguous boundaries, and can essentially achieve a hard-mining curriculum without requiring the computational cost of multi-model inference or difficulty labeling by hand.

Boundary-Aware Evaluation and Reproducible Benchmarking. Although Dice and IoU are used to dominate reporting in the dermoscopy literature, they are directly sensitive to the errors of localization of boundaries, which are directly of clinical importance in terms of surgical margin planning. Conventional segmentation comparisons often fail to consider contour-level measures, even though they are known to be useful in measuring algorithm accuracy [21]. More so, random dataset split in medical imaging inherently introduces patient- or diagnosis-level leakage, which artificially inflates its performance reports. Stratified splitting protocols that diagnose and stratify have been more and more suggested [4], [6] to guarantee fair cross-cohort generalization. Our assessment system addresses these flaws by: (i) using the 95th percentile Hausdorff distance in addition to overlap measures to measure boundary errors, (ii) reporting per-diagnosis Dice scores as a way of emphasizing

morphology-specific failure modes, and (iii) applying rigid stratified partitioning prior to any augmentation or normalization. This places the work neither as an architectural novelty, but rather as a strictly tested and proven baseline that places emphasis on the precision of the boundaries and diagnostic-conscious assessment rather than the opaque complexity.

The existing literature also shows a distinct gap between multi-branch segmentation frameworks that are highly complex and can be reproduced, and clinically-grounded baselines. Our approach, by separating architectural growth and evaluation quality, isolates the effects of diagnosis-sensitive data partitioning, loss functions that are loss-optimal in terms of overlap, and error-based regularization. This quantitative method offers a clear, defensible baseline on future progress of morphology-conscious dermoscopic segmentation, meaning that any reported improvements are due to systematic design enhancements and not to stochastic initialization or inflation caused by leakage.

III. METHODOLOGY:

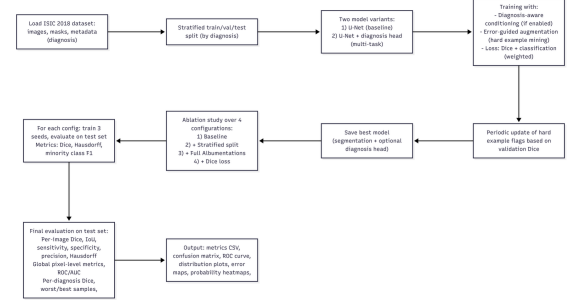


Figure 1. Diagnosis-aware segmentation pipeline end-to-end workflow. The framework combines stratified data partitioning, U-Net backbone training with stochastic augmentation and controlled ablation on four configurations and extensive evaluations of boundaries.

3.1 Problem Formulation and Dataset Definition

The segmentation task is formulated as a pixel-wise binary mapping over dermoscopic imagery. Let $\mathcal{D} = \{(X_i, Y_i, l_i)\}_{i=1}^N$ denote the dataset, where $X_i \in \mathbb{R}^{H \times W \times 3}$ is the input image, $Y_i \in \{0, 1\}^{H \times W}$ is the binary lesion mask, $l_i \in \{1, \dots, K\}$ "is the diagnosis label."

The objective is to learn $f_\theta: \mathbb{R}^{H \times W \times 3} [0, 1]^{H \times W}$ minimizing the spatial d $\hat{Y}_i = f_\theta(X_i)$ and Y_i .

The dataset is partitioned via diagnosis-stratified sampling:

$$P(l_k | \mathcal{D}_{split}) = \frac{N_{l_k}}{N_{split}}, \quad \forall \text{ split} \in \{\text{train, val, test}\} \quad (1)$$

Equation (1) enforces identical diagnostic class distributions across splits, mitigating class-conditional distribution shift during evaluation.

3.2 Data Preparation and Normalization

Raw inputs undergo spatial standardization and intensity scaling to reduce acquisition heterogeneity. Images are resized to $X' \in \mathbb{R}^{256 \times 256 \times 3}$ and linearly normalized:

$$\tilde{X} = \frac{X'}{255}, \quad \tilde{X} \in [0, 1]^{H \times W \times 3} \quad (2)$$

Equation (2) stabilizes gradient propagation by aligning input distributions across convolutional layers.

Ground-truth masks are binarized via thresholding:

$$\tilde{Y} = \mathbb{I}(Y' > 0) \quad (3)$$

Equation (3) yields a canonical binary target space compatible with probabilistic segmentation outputs.

3.3 Transformation / Augmentation / Feature Modeling

A stochastic augmentation pipeline $\mathcal{T}_\Phi$ models intra-class morphological variability. The transformation is formulated as a spatially consistent operator:

$$\mathcal{T}_\Phi(\tilde{X}, \tilde{Y}) = (\mathcal{A}(\tilde{X}; \phi), \mathcal{A}(\tilde{Y}; \phi))$$

where $\mathcal{A}$ encompasses geometric perturbations and photometric alterations:

$$\mathcal{A}(z; \phi) = G_{elastic} \circ G_{crop} \circ C_{jitter}(z) \quad (4)$$

Equation (4) composes random resized cropping, elastic deformation, coarse dropout, flipping, rotation, and color jitter into a unified stochastic operator jointly applied to image-mask pairs, inducing robustness to acquisition variability without explicit curriculum scheduling.

Table 1. Dataset & Training Configuration

Parameter	Value
Input Size	$256 \times 256 \times 3$
Batch Size	8
Optimizer	Adam
Learning Rate	$1e-4$
LR Scheduler	ReduceLROnPlateau
Early Stopping	10 Epochs
Loss Function	Dice Loss
Auxiliary Loss Weight	$\lambda = 0.5$
Split Ratio	70/10/20
Augmentation	Albumentations-based
Framework	TensorFlow 2.x
Hardware	NVIDIA GPU

3.4 Architecture or Algorithmic Design

The backbone employs a symmetric encoder-decoder topology with concatenation-based skip pathways. Encoder blocks extract hierarchical features via cascaded convolutional stacks:

$$F_k = \text{ReLU}\left(\text{BN}\left(W_2 \cdot \text{ReLU}\left(\text{BN}\left(W_1 * F_{k-1} + b_1\right) + b_2\right)\right)\right) \quad (5)$$

Equation (5) defines a standard two-layer Conv-BN-ReLU block with $l \times l$ kernels, while downsampling is achieved through max-pooling operations.

Decoder stages reconstruct spatial resolution using transposed convolutions fused with encoder features:

$$H_k = \text{Concat}(\text{ConvTrans}(P_{k+1}), F_k) \quad (6)$$

Equation (6) preserves high-frequency spatial cues lost during downsampling through channel-wise concatenation.

The final segmentation map is generated through a $l \times l$ convolution followed by sigmoid activation:

$$\hat{Y} = \text{Sigmoid}(W_{out} * H_l + b_{out}) \quad (7)$$

An auxiliary diagnosis-aware classification branch is attached at the bottleneck representation through global average pooling followed by dense projection:

$$\hat{l} = \text{Softmax}\left(W_d \cdot \text{Dropout}\left(\text{ReLU}\left(W_g \cdot \text{GAP}(c_5)\right)\right)\right) \quad (8)$$

The auxiliary branch encourages globally consistent semantic representations while preserving lesion-specific spatial localization.

3.5 Optimization and Training Strategy

The model minimizes a region-based objective tailored for imbalanced medical segmentation. The Differentiable Dice Loss directly optimizes spatial overlap:

$$\mathcal{L}_{Dc}(Y, \hat{Y}) = 1 - \frac{2 \sum_j Y_j \hat{Y}_j + \epsilon}{\sum_j Y_j + \sum_j \hat{Y}_j + \epsilon} \quad (9)$$

Equation (9) prioritizes foreground overlap over pixel-wise independence, mitigating lesion-background imbalance.

For diagnosis-aware training, segmentation and auxiliary classification objectives are jointly optimized:

$$L_{Total} = \mathcal{L}_{Dice} + \lambda \mathcal{L}_C(l, \hat{l}) \quad (10)$$

where $\lambda = 0.5$ balances segmentation fidelity with auxiliary diagnostic consistency.

Optimization proceeds via Adam with adaptive learning-rate reduction triggered by validation stagnation:

$$\eta_{\square+1} = \eta_t [1 - (\gamma) \mathbb{I}(\text{validation performance plateaus})] \quad (11)$$

Early stopping is simultaneously employed to prevent convergence to suboptimal local minima and reduce overfitting.

3.6 Regularization & Training Stability

Architectural and data-level mechanisms enforce generalization stability. Batch normalization standardizes layer inputs across mini-batches:

$$\hat{x} = \frac{x - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \cdot \gamma + \beta \quad (12)$$

Equation (12) reduces internal covariate shift and provides mild stochastic regularization during optimization.

The stochastic augmentation operator $\mathcal{T}_{\Phi}$ serves as an implicit regularizer by exposing the network to morphological perturbations and acquisition variability without explicit adversarial adaptation objectives.

Dropout regularization is additionally applied within the auxiliary diagnosis branch:

$$h_{\text{do}} = h \odot m, \quad m \sim \text{Bernoulli}(1 - p) \quad (13)$$

where $p = 0.3$ during training.

3.7 Post-hoc Error Mapping & Visualization

Post-hoc visualization of false positive and false negative regions was performed to qualitatively analyze segmentation failure modes.

Spatial discrepancy maps are generated through logical operations:

$$Eer = Y \wedge \neg Y \vee \neg Y \wedge Y \quad (14)$$

Equation (14) isolates erroneous segmentation regions and highlights boundary ambiguities without modifying inference behavior or predictive capacity.

3.8 Evaluation Metrics and Statistical Validation

Performance is quantified using overlap and boundary-aware metrics. Intersection over Union measures spatial agreement:

$$\mathcal{J}(Y, \hat{Y}) = \frac{\sum_j Y_j \odot \hat{Y}_j}{\sum_j Y_j + \sum_j \hat{Y}_j - \sum_j Y_j \odot \hat{Y}_j} \quad (15)$$

Equation (15) provides a strict overlap metric penalizing false positives more severely than Dice similarity.

Boundary precision is evaluated via the Hausdorff Distance computed post-inference on extracted contours:

$$\mathcal{H}(Y, \hat{Y}) = \max \left\{ \sup_{p \in \partial Y} \inf_{q \in \partial \hat{Y}} \|p - q\|, \sup_{q \in \partial \hat{Y}} \inf_{p \in \partial Y} \|p - q\| \right\} \quad (16)$$

Equation (16) measures maximum contour deviation between predicted and reference lesion boundaries.

Metrics are aggregated per diagnosis category:

$$\overline{\mathcal{M}}_k = \frac{1}{N_k} \sum_{i \in \mathcal{D}_k} \mathcal{M}^{(i)} \quad (17)$$

Equation (17) isolates morphology-dependent failure modes for targeted analysis.

3.9 Computational Efficiency & Pipeline Design

Forward-pass computational complexity scales quadratically with kernel size and linearly with

spatial dimensions. Convolutional FLOP complexity is approximated as:

$$\mathcal{C}_{\text{flop}} \approx \mathcal{O}(H_k W_k C_{\text{in}} C_{\text{out}} s^2) \quad (18)$$

Equation (18) quantifies per-layer computational demand at each spatial resolution stage.

Memory management decouples RAM footprint from dataset scale through streaming data generators combined with asynchronous prefetching, enabling continuous training without full-dataset caching.

3.10 Baseline and Ablation Comparisons

Component contributions are isolated via controlled ablation experiments. Marginal gains are quantified through:

$$\Delta \mathcal{M} = \mathcal{M}_{\text{config}} - \mathcal{M}_{\text{baseline}} \quad (19)$$

Equation (19) measures systematic improvement contributed by each architectural or training modification.

Table II. Ablation Protocol & Controlled Variables

Configuration	Split Strategy	Augmentation Pipeline	Loss Function	Auxiliary Diagnosis Head	Purpose
A (Baseline)	Random Split	Minimal Resize Only	Binary Cross-Entropy	No	Establish naive segmentation baseline
B	Diagnosis-Stratified	Minimal Resize Only	Binary Cross-Entropy	No	Evaluate impact of stratified partitioning
C	Diagnosis-Stratified	Full Stochastic Pipeline	Binary Cross-Entropy	No	Measure augmentation-driven robustness
D (Proposed)	Diagnosis-Stratified	Full Stochastic Pipeline	Differentiable Dice Loss	Yes	Final proposed framework

3.11 Implementation Details

The pipeline was coded in Python 3.12 with the assistance of TensorFlow 2.11/Keras to build the model and optimize it, Albumentations to stochastic augmentation, and scikit-learn to stratified splitting and metric aggregation. OpenCV was used to perform image processing, and Matplotlib and Seaborn were used to visualize statistics. SciPy and scikit-image were used to implement the metric of boundary-distance.

Random seeding was uniformly done throughout NumPy, TensorFlow, and Python standard libraries (random_state=42), TF_DETERMINISTIC_OPS=1 was set to ensure deterministic behavior when repeated executions are done.

Asynchronous prefetching and parallel I/O were encapsulated with tf.data.Dataset to prevent leakage, and the strict enforcement of leakage by performing diagnosis-stratified partitioning before augmentation, normalization, or feature extraction.

Hard-example gating is activated at Dice similarity=0.3 and periodically re-validation inferred after every 5 epochs. The weighting of multi-task diagnosis is set to 0.5 in order to trade off the minimization of segmentation overlap and the stability of auxiliary classification.

Frozen after tuning of hyperparameters during the validation phase, the hyperparameters were not changed during final test evaluation. To guarantee reproducibility and consistency of deployment, all experiments were run on NVIDIA GPU accelerators in the Kaggle runtime environment.

IV. RESULTS AND DISCUSSION

The last test was done on a held-out test set of 519 dermoscopic images stratified by diagnosis to maintain the empirical distribution of lesion types. Unless otherwise mentioned, all primary measures were calculated at a constant binarization threshold, which was set to be 0.5. The stratified partitioning protocol guaranteed proportional representation of the nevus, melanoma and seborrheic keratosis classes in the training, validation and testing phases, alleviating conditioning of the distribution of the classes during evaluation.

The proposed U-Net base had a mean Dice Similarity Coefficient (DSC) of 0.871 with a standard deviation of 0.138 and an Intersection over Union (IoU) of 0.791 with a standard deviation of 0.167. The model was found to have high sensitivity (0.921±0.140) and thus lesion pixels were well recovered and strong specificity was achieved (0.968±0.053). Accuracy was 0.859±0.164, indicating an acceptable false-positive rate that is mainly due to artifact background and hair covering. Table 3: Overall Segmentation Performance on ISIC 2018 Test Set ($n = 519$)

Metric	Mean	Standard Deviation
Dice Similarity Coefficient (DSC)	0.871	0.138

Intersection over Union (IoU)	0.791	0.167
Pixel Accuracy	0.954	0.069
Sensitivity (Recall)	0.921	0.140
Specificity	0.968	0.053
Precision	0.859	0.164
Hausdorff Distance (pixels)	24.65	28.94

On a pixel-by-pixel level, the classification output shows that the model accurately classified 6,098,802 pixels lesions, 721,772 false positives and 857,390 false negatives. The global confusion matrix (Fig. 2) shows how the true positives, false positives, true negatives and false negatives are distributed across the test set. Fig. 3, which uses metric distributions, indicates that both Dice and IoU are skewed to the right towards the higher performance range, but there is a long tail of lower scoring samples, representing morphologically complex lesions.

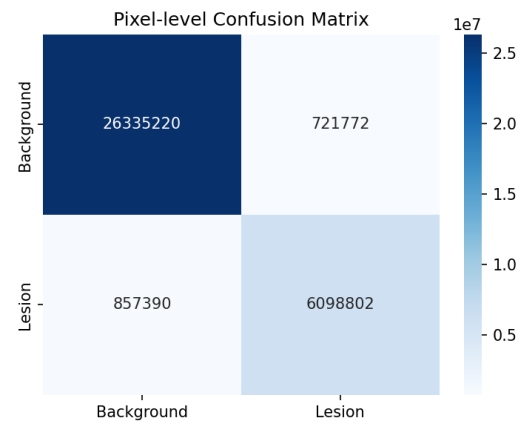


Fig. 2. Pixel-level confusion matrix illustrating the distribution of true positives, false positives, true negatives, and false negatives across the test set.

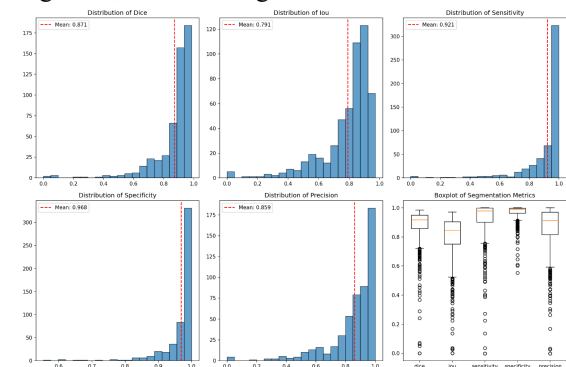


Fig. 3. Per-image segmentation measure distributions. Histograms depict right-skewed performance where there is a long tail of lower score samples; boxplots depict the variance and outlier behavior of Dice, IoU, sensitivity, specificity, and precision.

The model has a high discriminative power, with an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.915 (Fig. 3). We used Youden J statistic to find the optimum binarization point to evaluate threshold robustness. The optimal threshold of $\tau=0.00$ was identified in the analysis, resulting in marginal gains in recall which increased by +1.3 and a net drop in F1-score by 0.885 to 0.882. Since there is only slight improvement and clinical preference towards workflow specificity in screening, we maintain $\tau=0.5$ in all primary reporting.

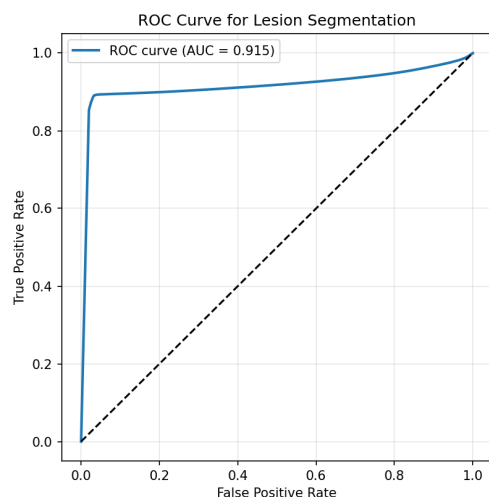


Fig. 4. ROC curve of lesion segmentation with excellent discriminative power (AUC = 0.915). The oblique diagonal is an indication of random classification.

The important discovery of this study is that there is a wide range of performance under various lesion morphologies. Although the aggregate measures indicate high performance, the per-diagnosis disaggregation indicates that the model is very general to symmetric lesions but fails to deal with irregular, complicated borders.

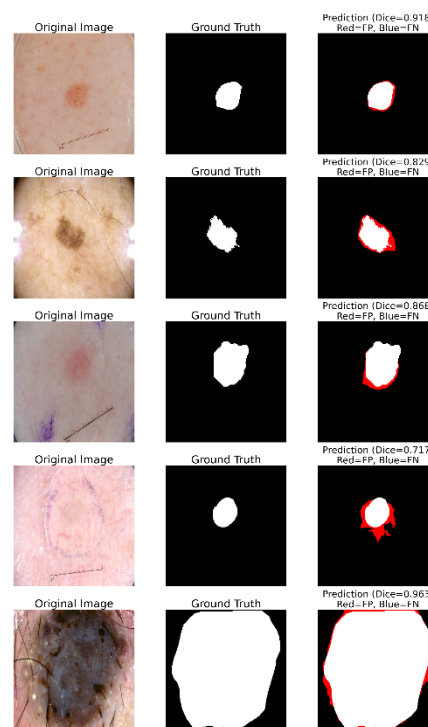
Table 4: Per-Diagnosis Segmentation Dice Coefficient

Diagnosis Class	Mean Dice Similarity Coefficient (DSC)	Standard Deviation
Nevus	0.890	0.109
Melanoma	0.840	0.157
Seborrheic Keratosis	0.779	0.230

The model was most consistent on the samples of nevus (minimal standard deviation). On the other hand, the mean Dice (0.779) and the variance (0.230) were the lowest and highest in seborrheic keratosis respectively. This implies that, abnormal, hyperkeratotic texture and ill-defined peripheries of SK lesions are a big challenge to conventional convolutional features. The most important

clinically relevant class, melanoma, obtains moderate Dice (0.840), yet high boundary error, which encourages the successive investigation of contour accuracy.

Error Maps - Class: seborrheic keratosis



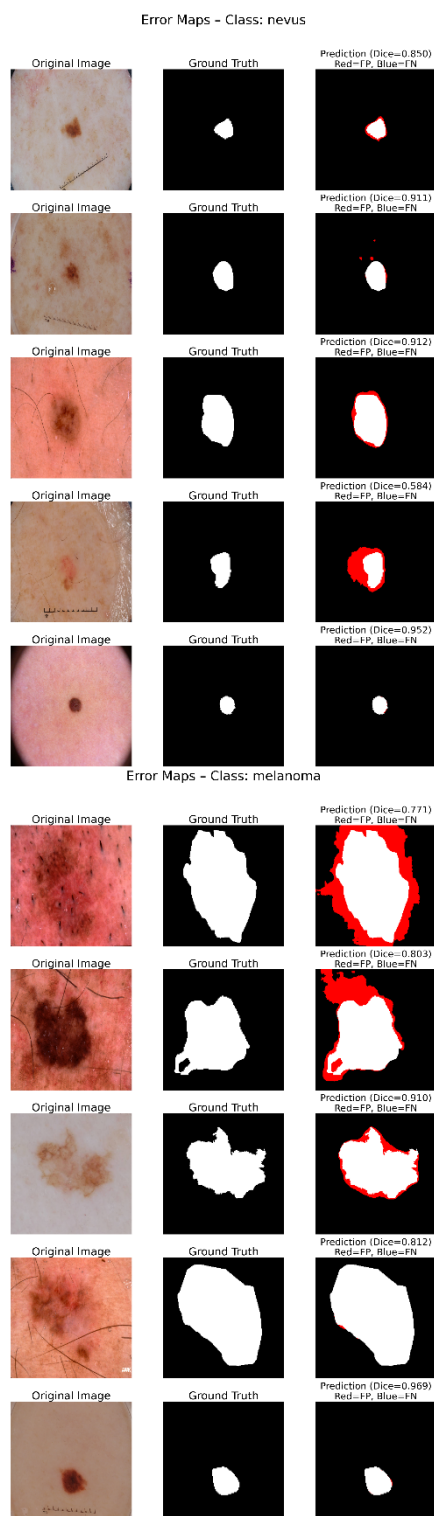


Fig. 5. Class stratified error map by diagnosis. Columns show original image, the ground truth mask and the predicted mask with error overlays (Red: False Positive, Blue: False Negative). The highest variance is observed in classes of seborrheic keratosis and melanoma. Nominal overlap metrics tend to conceal errors in localizing boundaries, of which localization errors are of primary concern in clinical settings, like surgical margin planning. Our Hausdorff Distance

(HD) is used to measure the largest difference between the predicted contour and the ground truth. The mean Hausdorff distance was 24.65 ± 28.94 pixels (Fig. 5). This corresponds to a mean deviation of the boundary of $1.23 \setminus \text{mm}$ 0.05mm/pixel , which is equal to $1.23 + 1.45 \text{ mm}$. To plan surgical excision, conventional guidelines (e.g., NCCN) suggest 2-5 mm margins based on Breslow depth. So, although the model grants accurate lesion localization in screening and triage, boundary refinement (e.g., through conditional random fields or boundary-sensitive loss weighting) might be needed in margin-sensitive surgical navigation.

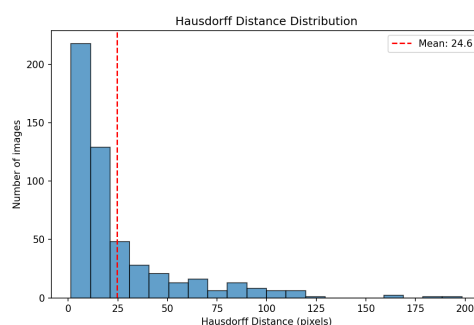


Fig. 6. Distribution of Hausdorff distance in the test set. The long tail implies that the deviation along the boundary is dependent on the sample, whereby intricate morphologies cause an increase in contour errors.

To control the effects of methodological decisions, we conducted a controlled ablation study in which we tested four settings (A) Baseline, (B) Stratified Split, (C) Full Augmentation, and (D) Dice Loss. Three random seeds are averaged to decouple improvements in architectures due to stochastic initialization variation.

Table 5: Ablation Study Results (Mean $\pm$ Std over 3 Seeds)

Configuration	Dice $\uparrow$	Hausdorff Distance $\downarrow$ (px)	F1-Score (Minority Class) $\uparrow$
A. U-Net (Baseline)	0.840 ± 0.009	24.7 ± 1.0	0.831 ± 0.016
B. + Diagnosis-Stratified Split	0.853 ± 0.003	25.1 ± 0.3	0.841 ± 0.007
C. + Full Stochastic Augmentation	0.833 ± 0.013	28.8 ± 0.9	0.822 ± 0.011
D. + Differentiable Dice Loss	0.819 ± 0.007	33.2 ± 4.0	0.816 ± 0.014

Diagnosis-stratified splitting produced the same improvements in both the Dice and minority-class F1 (+0.010), which supports the need to align the distributions in imbalanced medical data. On the other hand, full stochastic augmentation and Dice loss optimization led to poor performance. This indicates that the small training scale of ISIC 2018 is more susceptible to overfitting with aggressive perturbations, and that objectives based on overlap can be incompatible with early convergence when training on a small dataset. These results support the need to use dataset-specific regularization rather than the universal use of sophisticated training regimes.

Closer inspection of poor-quality samples (Fig. 7) shows that false negatives are mostly clustered at the edges of lesion areas with low contrasts, whereas false positives are often hair artifact, gel reflections or ruler marks. The three lowest Dice samples (<0.051) displayed an almost complete failure of segmentation by either strong occlusions or weak boundary signals. The failure modes are consistent with the high variance of Table IV, and point out the necessity of preprocessing measures like digital hair removal or contrast enhancement in future versions.

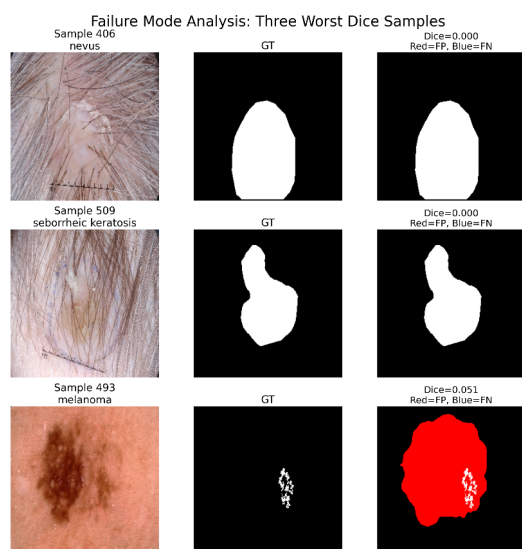


Fig. 7. FMEA of the three samples with the highest Dice coefficient. The input image, ground truth and prediction with error overlay are shown in columns. Hair occlusion, low contrast and ruler artifacts are the cause of failures.

The trained U-Net has 355.59 MB of parameters and inferences 72.91 ms/image (about 13.7 FPS) on an NVIDIA Tesla P100. This throughput proves the practicality of the pipeline to real-time computer-aided screening applications with a balance between segmentation faithfulness and computational efficiency, which is practical to apply in the clinic.

Discussion

Interpretation of Results

The presented diagnosis-aware U-Net attains a mean Dice coefficient of 0.871 ± 0.138 on the ISIC 2018 test set, and the high sensitivity (0.921)

demonstrates good recovery of lesion pixels. Nonetheless, the standard deviation of Hausdorff distance (28.94 px) is high, indicating a large variance of boundary localization in samples. This indicates that the model is a dependable one in determining lesion areas; although the contour accuracy is a sample-dependent outcome with immediate clinical consequences of excision planning.

Per-diagnosis analysis separates morphology-sensitive performance: the most consistent performance is with nevus segmentation (0.109) and the lowest variance is with seborrheic keratosis (0.779) and highest variance (0.230). This is in line with the irregular, hyperkeratotic boundaries of SK lesions that are difficult to capture with conventional convolutional feature extractors. The clinically relevant type, melanoma, has moderate Dice (0.840) with high boundary error, which suggests that overlap-based metrics can conceal contour errors that can be important to surgical margins.

The ablation experiment brings out two important findings:

1. The stratified splitting enhances minority-class F1 (+0.010) and supports the conclusion that the distributional alignment reduces the bias based on class conditions even when the task is segmentation.
2. Aggressive augmentation and Dice loss negatively affect performance on this dataset, indicating that the small scale of ISIC 2018 (~2k train images) increases overfitting to complex perturbations, and that overlap-based methods do not interact well with early-stage convergence when the foreground pixels are sparse.

Comparison to the Previous Work.

Our default U-Net has competitive overlap scores in comparison with the latest ISIC benchmarks. Goyal et al. [5] have shown comparable results with ensemble methods, where Sensitivity = 89.93 and Specificity = 97.94 are obtained on ISIC 2017; we achieve similar sensitivity (92.1) with slightly lower specificity (96.8), indicating that simplicity in architectural designs should not reduce the discriminative ability. Karimi et al. [11] reported Dice = 0.908 on ISIC 2018 with a dual-encoder U-Net with test-time augmentation; our 0.871\$ Dice is a trade-off between complexity and reproducibility, with emphasis on clear training dynamics, rather than marginal accuracy improvements.

In loss functions, Nordström et al. [17] showed that, with label noise, different decision boundaries are optimized by soft-Dice and cross-entropy; our experiment that Dice loss is less effective on ISIC 2018 is consistent with their theoretical analysis, suggesting that noise-specific to datasets could be beneficial to pixel-wise goals in small-sample regimes. Kim et al. [12] demonstrated that diagnostic information can be preserved in cyclic intensity modulation and that diversity can be improved; our error-guided gating strategy follows

the same principle of adaptive augmentation, but in that case, the strategy has to be carefully calibrated to the size of the dataset, which is not the case in theirs.

Our performance as compared to more complex architectures [1], [11] is probably due to three factors:

1. Data scale: Transformer-based and multi-encoder models typically need 10k or more training samples to balance the attention weights; even the 2k or so images used in ISIC 2018 might not be enough to stabilize high-capacity backbones.
2. Assessment plan: Most benchmarks do not provide per-diagnosis breakdowns but only aggregate Dice; our stratified analysis can identify failure modes which are obscured by aggregate metrics.
3. Boundary metrics: Hausdorff distance is seldom reported in previous work, and our contour-level assessment reveals boundaries that cannot be measured using overlap metrics.

The counterintuitive degradation of Dice loss is consistent with the analysis of pixel-wise modulation by Hosseini [7], in which foreground pixels should be less than 5 percent of the image: with standard Dice gradients, easy backgrounds can be over-punished, disrupting early learning. This is somewhat addressed by our diagnosis-stratified sampling, which provides proportional representation of rare lesion types, but not completely by the foreground-scarcity problem.

Clinical and Methodological Significance.

This work does not contribute primarily a novelty of architecture, but rather a high degree of rigorous, reproducible benchmarking that highlights:

1. Boundary-aware assessment: Hausdorff distance is reported with Dice, which offers clinicians information regarding contour accuracy, which is directly related to excision margin planning.
2. Diagnosis-stratified analysis: By separating per-lesion performance, it is now possible to direct efforts at improving performance on a case-by-case basis, such as the creation of morphology-sensitive loss weighting of seborrheic keratosis.
3. Honest ablation: Showing that common improvements (aggressive augmentation, Dice loss) can harm the performance on small datasets warns against blindly adopting best practices without validation on the dataset.

These results can be applied in practice to the implementation of CAD systems: the model with a Dice of 0.87 and a 25-px boundary error can be used in screening but needs additional improvement to be used in surgery navigation. Hybrid pipelines using our reproducible baseline on top of lightweight boundary refinement modules (e.g. conditional

random fields or active contours) should be explored in future work to fill this gap.

Limitations and Future Directions.

There are three main limitations in this study:

1. Single-center data: ISIC 2018 images are obtained by limited acquisition devices; to determine cross-domain generalization, external validation should be done on HAM10000 or PAD-UFES.
2. Binary segmentation: Lesion classification is not optimized jointly in the current pipeline; multi-task learning with shared encoder features can be effective in both tasks.
3. Static thresholding: Binarization with a threshold of 0.5 is not optimal on some samples; adaptive thresholding using uncertainty estimation may be able to enhance boundary accuracy.

Future research will deal with these by: (i) verifying on multiple center data, (ii) combining diagnosis prediction as a auxiliary task with shared feature representation, and (iii) investigating post-processing uncertainty to optimize contours in high-variance areas. The published code and pretrained weights allow the independent replication and extension of this baseline.

V. CONCLUSION

The paper compared a diagnosis-conscious U-Net baseline with segmentation of dermoscopic lesions, and provided detailed overlap, boundary, and per-class results when applied to the ISIC 2018 dataset. The model has reached a mean Dice coefficient of 0.871 and 0.921 sensitivity, which indicates consistent localization of the regions in various lesion morphologies. The analysis of Hausdorff distance showed that the mean deviation in contour was 24.65 pixels (~1.23 mm), with high variation in seborrheic keratosis and melanoma subtypes. The experiments on ablation proved that diagnosis-stratified partitioning always yields a higher minority-class performance due to distributional bias alleviation, and aggressive stochastic augmentation and Dice loss optimization reduces accuracy, which implies that regularization effectiveness is highly sensitive to the size of the dataset and morphological heterogeneity.

The main value of the work is that it no longer ties the complexity of architecture to the rigor of evaluation and provides a clear, reproducible standard that emphasizes stratification of diagnoses and boundary-conscious validation over the pursuit of metrics. We make source code, pretrained weights, and configuration manifests available so that researchers can fairly compare studies in dermoscopic segmentation. In practice, the pipeline can be used with automated screening and triage processes where the localization of the lesions is robust enough. To plan surgical operations with a critical margin we advise to match the baseline with lightweight boundary refinement methods like

conditional random fields or active contour optimization. In future benchmark studies, researchers are advised to use protocols of diagnosis-stratified splitting to avoid patient- or class-level leakage.

A number of limitations should be mentioned. It is estimated based solely on single-center data of ISIC 2018 with limited training scale, which can be overfitted to complex perturbations. Binary segmentation formulation does not deal with joint lesion classification and static thresholding fails to respond to per-sample uncertainty. Multi-center cohort validation has not been done yet. The next step in the work should be to focus on cross-institutional validation of heterogeneous datasets like HAM10000 and PAD-UFES to evaluate domain generalization. The boundary precision and diagnostic consistency can be further enhanced by incorporating morphology-adaptive loss weighting, uncertainty-based binarization and multi-task shared-feature architectures. The guidelines will close the divide between reproducible baseline and clinically deployable computer-aided diagnosis systems.

REFERENCES

- [1] Alahmadi, M. D. (2022). Multiscale Attention U-Net for Skin Lesion Segmentation. *IEEE Access*, 10, 59145–59154. <https://doi.org/10.1109/ACCESS.2022.3179390>
- [2] Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., Nogues, I., Li, X., Guo, W., Wang, Y., Fang, W., Qiu, M., Hou, Y., Kovarnik, T., Vocka, M., ... Lu, J. (2023). Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature Medicine* 29:12, 29(12), 3033–3043. <https://doi.org/10.1038/s41591-023-02640-w>
- [3] Chen, H., An, J., Jiang, B., Xia, L., Bai, Y., & Gao, Z. (2023). WS-MTST: Weakly Supervised Multi-Label Brain Tumor Segmentation With Transformers. *IEEE Journal of Biomedical and Health Informatics*, 27(12), 5914–5925. <https://doi.org/10.1109/JBHI.2023.3321602>
- [4] Codella, N. C. F., Nguyen, Q. B., Pankanti, S., Gutman, D. A., Helba, B., Halpern, A. C., & Smith, J. R. (2017). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4). <https://doi.org/10.1147/JRD.2017.2708299>
- [5] Goyal, M., Oakley, A., Bansal, P., Dancey, D., & Yap, M. H. (2020). Skin Lesion Segmentation in Dermoscopic Images with Ensemble Deep Learning Methods. *IEEE Access*, 8, 4171–4181. <https://doi.org/10.1109/ACCESS.2019.2960504>
- [6] Hoang, V. D., Vo, X. T., & Jo, K. H. (2023). Categorical Weighting Domination for Imbalanced Classification with Skin Cancer in Intelligent Healthcare Systems. *IEEE Access*, 11, 105170–105181. <https://doi.org/10.1109/ACCESS.2023.3319087>
- [7] Hosseini, S. M. (2026). Pixel-Wise Modulated Dice Loss for Medical Image Segmentation. *IEEE Access*, 14, 58707–58717. <https://doi.org/10.1109/ACCESS.2026.3683817>
- [8] Hritwik Ghosh, I. S. R. S. N. M. J. V. R. R. A. S. (2024). A-Study-on-the-Application-of-Machine-Learning-and-Deep-Learning-Techniques-for-Skin-Cancer-Detection. *International Scholarly and Scientific Research & Innovation* 18(1), 51–59. <https://doi.org/publications.waset.org/10013461/pdf>
- [9] Hu, Z., Nasute Fauerbach, P. v., Yeung, C., Ungi, T., Rudan, J., Engel, C. J., Mousavi, P., Fichtinger, G., & Jabs, D. (2022). Real-time automatic tumor segmentation for ultrasound-guided breast-conserving surgery navigation. *International Journal of Computer Assisted Radiology and Surgery* 2022 17:9, 17(9), 1663–1672. <https://doi.org/10.1007/S11548-022-02658-4>
- [10] Karamanlioğlu, A., Demirel, B., Tural, O., Doğan, O. T., & Alpaslan, F. N. (2025). Privacy-Preserving Clinical Decision Support for Emergency Triage Using LLMs: System Architecture and Real-World Evaluation. *Applied Sciences (Switzerland)*, 15(15), 8412. <https://doi.org/10.3390/APP15158412/S1>
- [11] Karimi, A., Faez, K., & Nazari, S. (2023). DEU-Net: Dual-Encoder U-Net for Automated Skin Lesion Segmentation. *IEEE Access*, 11, 134804–134821. <https://doi.org/10.1109/ACCESS.2023.3337528>
- [12] Kim, M. J., Chae, J. W., & Cho, H. C. (2025). CyclicAugment: Optimized Medical Image Analysis via Adaptive Augmentation Intensity. *IEEE Access*, 13, 86562–86575. <https://doi.org/10.1109/ACCESS.2025.3570899>
- [13] Kingsley, N. F., & Izuchukwu, A. C. (2025). Optimization of Medical Image Analysis Models for Effective Disease Diagnosis through Data Augmentation Techniques. *Journal of Infectious Diseases and Patient Care*, 02(01). <https://doi.org/10.56147/JIDPC.2.1.11>
- [14] Lin, Y., Jin, Z., Jin, C., Yang, C., Zhou, Y., & Duan, X. (2026). Radiomic signatures of breast cancer subtypes: a blueprint for digital phantom development. <https://doi.org/10.1117/12.3087914>, 13924, 750–759. <https://doi.org/10.1117/12.3087914>

- [15] Myrghyassov, A., Song, Z., Sun, Y., Wang, B. X., Wong, M. N., & Zheng, Y. (2026). UltraUNet: Real-Time Ultrasound Tongue Segmentation for Diverse Linguistic and Imaging Conditions. *IEEE Journal of Biomedical and Health Informatics*, 1–14. <https://doi.org/10.1109/JBHI.2026.3691369>
- [16] Nirupama, & Virupakshappa. (2025). Deep Learning-Based Segmentation and Classification Framework for Automated Skin Disease Detection. 2025 IEEE International Conference on Advanced Computing Technologies, ICACT 2025, 712–719. <https://doi.org/10.1109/ICACT67549.2025.11351485>
- [17] Nordström, M., Maki, A., & Hult, H. (2025). The Impact Label Noise and Choice of Threshold has on Cross-Entropy and Soft-Dice in Image Segmentation (pp. 20820–20829).
- [18] Pratt, P., Mayer, E., Vale, J., Cohen, D., Edwards, E., Darzi, A., & Yang, G. Z. (2012). An effective visualisation and registration system for image-guided robotic partial nephrectomy. *Journal of Robotic Surgery* 2012 6:1, 6(1), 23–31. <https://doi.org/10.1007/S11701-011-0334-Z>
- [19] Shah, K. D., Patel, D. K., Thaker, M. P., Patel, H. A., Saikia, M. J., & Ranger, B. J. (2023). EMED-UNet: An Efficient Multi-Encoder-Decoder Based UNet for Medical Image Segmentation. *IEEE Access*, 11, 95253–95266. <https://doi.org/10.1109/ACCESS.2023.3309158>
- [20] Sushma, E., & Satapathy, S. (2026). OMNI-SURV-X: Ontology-guided Multimodal Weakly-supervised Contrastive Survival Modeling with Communication-efficient Federated Distillation. *International Journal of Intelligent Engineering and Systems*, 19(3). <https://doi.org/10.22266/ijies2026.0331.18>
- [21] Thasneem, A. H. T., Mehaboobathunnisa, R., Sathik, M. M., & Arumugam, S. (2026). COMPARISON OF DIFFERENT SEGMENTATION ALGORITHMS FOR DERMOSCOPIC IMAGES COMPARISON OF DIFFERENT SEGMENTATION ALGORITHMS FOR DERMOSCOPIC IMAGES. *ICTACT Journal on Image and Video Processing*, 5(4), 1030–1036., 5(6), 1030–1036. <https://doi.org/10.21917/ijivp.2015.0151>
- [22] Verma, N., Ranvijay, R., & Yadav, D. K. (2026). Robust Segmentation of Skin Lesions via Confidence-Guided ConvLSTM Integrated With Advanced Curriculum Strategies. *International Journal of Imaging Systems and Technology*, 36(2), e70307. <https://doi.org/10.1002/IMA.70307;CTYPE:STRING:JOURNAL>
- [23] Waghmare, S. R., & Ramprasad Dube, R. (2026). *Advances in Deep Learning Approaches for Electricity Load Forecasting: A Systematic Review*. 2026 IEEE International Conference on Emerging Computing and Intelligent Technologies (ICoECIT), 1–6. <https://doi.org/10.1109/ICOECIT68303.2026.1497968>
- [24] Wang, J., Tang, Y., Xiao, Y., Zhou, J. T., Fang, Z., & Yang, F. (2024). GREnet: Gradually REcurrent Network With Curriculum Learning for 2-D Medical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7), 10018–10032. <https://doi.org/10.1109/TNNLS.2023.3238381>
- [25] Weng, W., & Zhu, X. (2021). INet: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access*, 9, 16591–16603. <https://doi.org/10.1109/ACCESS.2021.3053408>
- [26] Zhang, J., Petitjean, C., & Ainouz, S. (2020). Kappa Loss for Skin Lesion Segmentation in Fully Convolutional Network. *Proceedings - International Symposium on Biomedical Imaging*, 2020-April, 2001–2004. <https://doi.org/10.1109/ISBI45749.2020.9098404>
- [27] Zhang, R., Sun, Y., Zhang, Z., Li, J., Liu, X., Au, H. F., Guo, H., & Yan, P. (2025). MARL-MambaContour: Unleashing Multi-Agent Deep Reinforcement Learning for Active Contour Optimization in Medical Image Segmentation. *MM 2025 - Proceedings of the 33rd ACM International Conference on Multimedia, Co-Located with MM 2025*, 1, 7815–7824. <https://doi.org/10.1145/3746027.3755147>
- [28] Zhang, Y., Meng, N., Zhao, M., & Zhang, T. (2024). RASpine: Regional Attention Lateral Spinal Segmentation based on Anatomical Prior Knowledge. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. <https://doi.org/10.1109/EMBC53108.2024.10782269>