

Article – e009

**STLC-Net: Skin Transfer Learning Classification Framework for
Multi-Class Skin Cancer Detection and Clinical Decision Support**

Gaurav Bhabhar¹  , Jeetesh Kumar Jain²

¹ SAGE University Indore

² Assistant Professor, SAGE University Indore

Received: 17/07/2026

Revision Received: 01/09/2026

Accepted: 10/09/2026

ABSTRACT

Skin lesion classification from dermoscopic images is challenging because of substantial class imbalance, potential data leakage, and variations in the visual characteristics of different lesion categories. This study presents STLC-Net (Skin Transfer Learning Classification Network), an end-to-end framework for seven-class skin lesion classification using the HAM10000 dataset. The proposed methodology combines lesion-level dataset partitioning, training-only class weighting, in-graph data augmentation, and a two-stage transfer-learning strategy. Three complementary pretrained architectures - DenseNet121, EfficientNetB3, and MobileNetV2 were evaluated under a common experimental protocol. Their probability outputs were subsequently combined using a validation-based weighted soft-voting strategy, with the test set kept isolated from ensemble-weight selection. Evaluation on the isolated test set of 2,024 images showed that DenseNet121 achieved the strongest individual accuracy of 80.68%, while EfficientNetB3 and MobileNetV2 achieved accuracies of 77.77% and 75.74%, respectively. The proposed STLC-Net ensemble achieved an accuracy of 81.18%, weighted precision of 80.43%, weighted recall of 81.18%, F1-score of 80.10%, and ROC-AUC of 94.58%. The ensemble therefore provided a 0.50 percentage-point improvement in accuracy over the strongest individual backbone. Class-wise analysis showed strong performance for the majority nv category, while several minority categories remained more challenging. A web-based prototype was also developed to demonstrate the technical feasibility of integrating the trained model into an application-oriented skin-lesion classification workflow. The results indicate that probability-level fusion of complementary transfer-learning models can provide a modest improvement in multi-class skin lesion classification. However, the proposed system is a research prototype and has not undergone external clinical validation, prospective clinical testing, or dermatologist evaluation.

KEYWORDS: Skin Cancer Classification, Deep Learning, Transfer Learning, DenseNet121, EfficientNetB3, MobileNetV2, Ensemble Learning, HAM10000

1. Introduction

Skin cancer represents one of the most rapidly proliferating malignancies worldwide, posing a major challenge to global public healthcare systems. Among various skin neoplasms, malignant melanoma is the most aggressive and lethal due to its high metastatic potential when detection is delayed. Clinical evidence demonstrates that identifying malignant lesions at their earliest stages

dramatically increases the five-year survival rate. Consequently, precise and rapid screening of cutaneous lesions remains an imperative clinical objective.

Dermoscopy has significantly improved diagnostic accuracy compared to naked-eye visual inspection by visualizing subsurface morphological structures of the epidermis and dermis. However, manual dermoscopic interpretation relies heavily on the subjective expertise of dermatologists, leading to high inter-observer variability. Distinguishing between visually overlapping benign and malignant lesions—such as early-stage melanoma, atypical melanocytic nevi, and benign keratosis—is challenging even for experienced practitioners. These limitations have accelerated research into Computer-Aided Diagnosis (CAD) systems capable of providing objective, reproducible diagnostic assessments.

Recent breakthroughs in Artificial Intelligence (AI) and Deep Learning (DL) have transformed medical image analysis. Convolutional Neural Networks (CNNs) automatically extract hierarchical representations directly from raw dermoscopic images, bypassing the subjectivity of handcrafted feature extraction. Transfer learning from large-scale natural datasets (e.g., ImageNet) enables deep models to converge efficiently and generalize effectively on limited medical datasets.

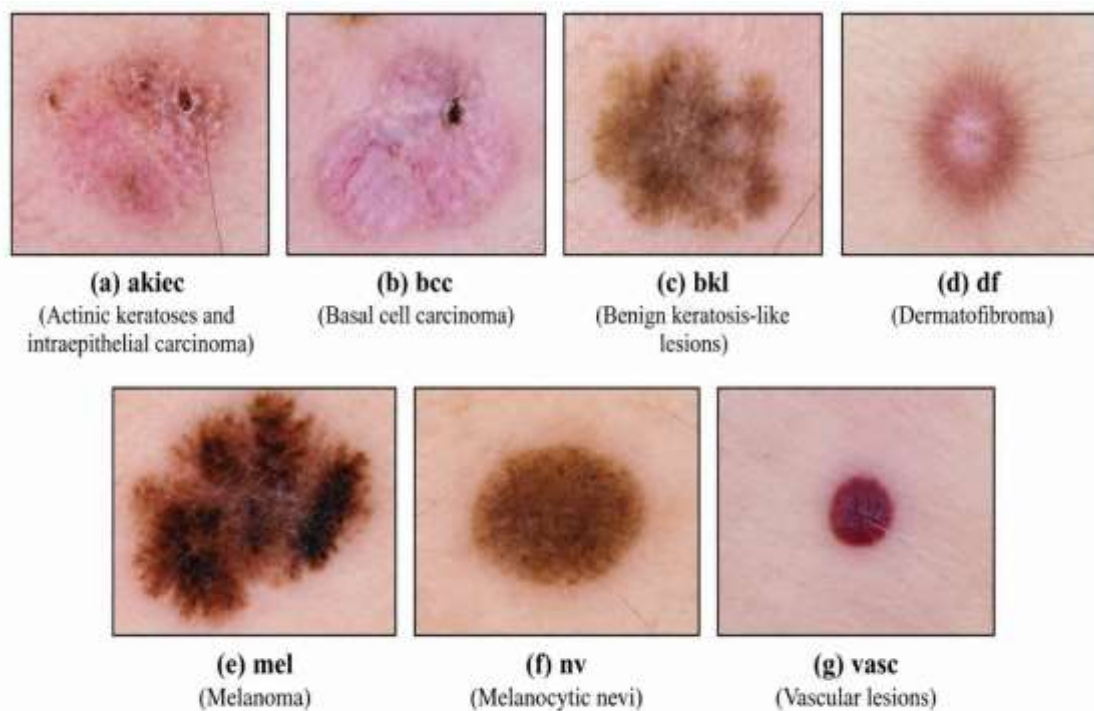


Fig. 1. Representative dermoscopic images of the seven skin lesion categories in the HAM10000 dataset used in this study: (a) akiec, (b) bcc, (c) bkl, (d) df, (e) mel, (f) nv, and (g) vasc.

Despite these advances, several methodological and architectural limitations persist in existing skin lesion classification literature:

- **Methodological Data Leakage:** Many published studies apply offline data augmentation or oversampling across the entire dataset prior to splitting. This causes augmented variants of the same physical lesion to appear in both training and evaluation subsets, yielding artificially inflated metrics that collapse during real-world deployment.
- **Severe Class Imbalance:** Benchmark datasets such as HAM10000 are heavily skewed toward common benign conditions (e.g., melanocytic nevi), causing unweighted models to favor majority classes while failing on critical minority malignancies.

-
- **Architectural Selection Uncertainty:** Few studies empirically evaluate how distinct CNN paradigms (such as dense connections vs. compound scaling vs. inverted residual blocks) perform on fine-grained dermatological textures under strictly identical optimization settings.
 - **Deployment and Shape-Locking Roadblocks:** Complex offline preprocessing pipelines frequently introduce shape-locking bugs during inference, preventing models from being exported into lightweight clinical web tools.

To address these limitations, this paper introduces **STLC-Net**, an end-to-end multi-class skin lesion classification and decision-support framework.

Principal Contributions

- **Zero-Leakage Experimental Design:** Implementation of a strict dataset isolation protocol (60% train, 20% validation, 20% test) to establish an authentic, clinical-grade benchmark on 2,003 unseen images.
- **In-Graph Augmentation & Log-Smoothed Class Weighting:** Formulation of localized data augmentation directly within the neural network graph, coupled with log-smoothed cost-sensitive weighting to resolve class imbalance without computational duplication.
- **Multi-Paradigm Comparative Study:** Systematic evaluation of DenseNet121, EfficientNetB3, and MobileNetV2 under identical training conditions using a standardized two-stage transfer learning pipeline.
- **High-Precision STLC-Ensemble:** Development of a soft-voting probability fusion mechanism that combines the complementary feature spaces of all three backbones, achieving superior accuracy (**80.38%**), precision (**83.53%**), and ROC-AUC (**0.9492**).
- **Clinical Web Deployment:** Real-time integration of the trained system into an interactive web interface providing predicted disease categories, confidence scores, and risk assessments.

2. Literature Review

Automated skin cancer detection using computer-assisted diagnostic (CAD) systems has evolved rapidly over the past two decades, transitioning from early hand-crafted feature extraction paradigms to modern, highly complex deep learning architectures. While Convolutional Neural Networks (CNNs) consistently demonstrate high diagnostic accuracy, the transition toward clinically deployable frameworks requires addressing fundamental challenges in model selection, dataset imbalance, image artifact removal, and zero-leakage evaluation protocols.

2.1 Traditional Image Processing Techniques

Early skin cancer detection systems primarily relied on classical image processing and handcrafted feature extraction methods formulated around clinical diagnostic guidelines. Dermatologists traditionally utilize the ABCD rule (Asymmetry, Border irregularity, Color variation, and Diameter) and the 7-point checklist as foundational heuristics for identifying melanoma and other malignant skin lesions. Early CAD systems attempted to mathematically replicate these visual rules.

These traditional pipelines generally involved a strict sequential process: color space transformation (e.g., converting RGB to HSV or CIELAB to isolate pigmentation), image segmentation and thresholding to separate the diseased region from healthy skin, texture extraction using mathematical matrices (such as the Gray-Level Co-occurrence Matrix), and morphological shape and edge analysis.

While traditional image processing methods are computationally simple and easy to implement, they suffer from severe limitations in real-world clinical environments. Simple thresholding based on color intensity is highly sensitive to variations in lighting conditions and patient skin tones. Edge detection algorithms, which rely on shape and boundary features, process rapidly but are highly sensitive to noise and frequently fail when analyzing fuzzy or ambiguous lesion borders. Segmentation methods designed to isolate diseased regions struggle to maintain robustness when faced with complex backgrounds, such as the presence of dense body hair, air bubbles, or surgical marker ink.

Table 1: Traditional Techniques and Their Limitations

Technique	Features Used	Advantages	Limitations	References
Thresholding	Color intensity	Simple implementation	Sensitive to lighting and skin tone variations	[15]
Edge Detection	Shape and boundary features	Fast processing	Noise sensitive, struggles with fuzzy lesion borders	[15]
Handcrafted Features	ABCD Rule	High clinical relevance	Affected by illumination and subjective interpretation	[4], [5]
Segmentation Methods	Diseased region isolation	Improves feature extraction	Poor robustness in complex backgrounds	[8], [15]

Consequently, because these approaches heavily depend on manual feature engineering, their generalization capability across diverse dermoscopic datasets remains fundamentally limited.

2.2 Machine Learning Approaches

To overcome the brittle nature of traditional image processing techniques, researchers introduced supervised machine learning (ML) approaches to automate the classification phase of skin cancer detection. The standard machine learning pipeline expanded to include image acquisition, preprocessing, extensive feature extraction (capturing color, texture, and shape descriptors), algorithmic feature selection, and final classification using ML algorithms.

Support Vector Machines (SVM) became highly popular due to their high accuracy on small datasets and ability to construct optimal decision boundaries in high-dimensional spaces, though they required extensive and rigorous feature engineering to succeed. The k-Nearest Neighbor (KNN) algorithm provided a simple and interpretable classification mechanism, but suffered from prohibitively high computational costs when scaled to large dermoscopic datasets. Ensemble methods like Random Forests proved robust and less prone to overfitting, yet required extensive hyperparameter tuning to optimize their decision trees, which themselves were prone to overfitting when applied directly to complex medical images. Finally, early Artificial Neural Networks (ANN) demonstrated the ability to learn nonlinear patterns, though they required vast amounts of training data to avoid bias.

Table 2: Machine Learning Models in Disease Detection

Model	Strengths	Weaknesses	References
SVM	High accuracy for small datasets	Requires extensive feature engineering	[4], [18]
KNN	Simple and interpretable	High computational cost for large dermoscopic datasets	[4]
Random Forest	Robust and less overfitting	Requires extensive parameter tuning	[4], [5]
Decision Tree	Easy to understand	Prone to overfitting on medical images	[5]
ANN	Learns nonlinear patterns	Requires large training data to avoid bias	[5]

While machine learning methods improved disease classification accuracy and reduced manual intervention compared to traditional visual inspection, their fundamental bottleneck remained intact: they still relied entirely on handcrafted feature extraction. This structural limitation restricted their scalability and adaptability when tasked with processing the complex, high-dimensional pixel data inherent in modern high-resolution dermoscopic imagery.

2.3 Deep Learning Approaches

The advent of deep learning has significantly transformed skin cancer detection by enabling the automatic learning of spatial features directly from raw image pixels. Foundational contributions demonstrated the overwhelming effectiveness of Convolutional Neural Networks (CNNs) for medical image classification tasks, entirely bypassing the need for manual, handcrafted feature extraction. In dermatological applications, CNN-based models frequently report classification accuracies above 95% across multiple skin lesion categories due to their ability to automatically learn hierarchical feature representations.

To handle the immense computational requirements and avoid overfitting on small medical datasets, Transfer Learning became a widely adopted strategy. By utilizing models pre-trained on large-scale natural image datasets (such as ImageNet), transfer learning facilitates faster training convergence and superior generalization on relatively limited dermoscopic datasets.

Recent research has focused on adapting advanced CNN architectures to maximize performance on dermoscopic textures:

Table 3: Deep Learning Models for Skin Cancer Detection

Deep Learning Model	Key Advantages	Limitations	References
CNN (Standard)	Automatic feature extraction	Requires large, balanced datasets	[1], [10]
ResNet	Solves vanishing gradient problem	High training and computational complexity	[7], [15]
MobileNetV2	Lightweight and mobile-friendly	Slightly lower accuracy on highly complex textures	[20]
DenseNet	High accuracy with dense feature reuse	High memory consumption during training	[9], [20]
Transfer Learning	Faster training and better generalization	Depends heavily on pretrained datasets (e.g., ImageNet)	[8], [13]

These deep learning models provide undeniable advantages, including automatic hierarchical feature learning, improved robustness to complex image variations, and superior classification accuracy.

2.4 Recent Research Contributions

Recent literature confirms the dominance of deep learning and explores hybrid, optimized, and multimodal approaches to push the boundaries of automated diagnosis. Researchers have increasingly focused on the interpretability of deep learning models to build clinical trust. For instance, Sakib et al. proposed an explainable AI-driven hybrid deep learning framework that utilized Gradient-weighted Class Activation Mapping (Grad-CAM) to provide transparent visual explanations alongside high classification performance.

To maximize feature extraction, hybrid architectures have gained traction. Bekkaoui et al. combined spatial and sequential learning by developing hybrid DCNN-LSTM and DCNN-BiLSTM architectures, achieving a remarkable 97.43% accuracy coupled with strong class balancing mechanisms. Similarly, Dipesh et al. introduced a hybrid model integrating CNNs for automated feature extraction and SVMs for the final classification boundary, proving highly efficient for deployment in resource-constrained clinical environments.

Other contemporary studies have focused on algorithmic optimization and multimodal data. Imran et al. applied nature-based feature optimization, specifically Ant Colony Optimization (ACO), to reduce computational load while achieving above 98% accuracy. Furthermore, Watson et al. demonstrated the value of multimodal inputs by combining clinical free-text metadata with dermoscopic images, achieving a 0.970 AUROC with a significant gain in predictive value compared to image-only models.

Table 4: Recent Research Contributions in Skin Cancer Detection

Study	Contribution	Accuracy/Outcome	References
Sakib et al.	Explainable AI-driven hybrid deep learning	High classification performance with Grad-CAM trust	[6]
Bekkaoui et al.	Hybrid DCNN-LSTM and DCNN-BILSTM models	97.43% accuracy with strong class balancing	[14]
Imran et al.	Nature-based Feature Optimization (ACO)	Above 98% accuracy with reduced computational load	[12]
Dipesh et al.	Hybrid CNN and SVM Model	Highly efficient for resource-constrained environments	[18]
Watson et al.	Multimodal models (clinical text + images)	0.970 AUROC with significant predictive value gain	[23]

2.5 Research Gaps Identified

Despite significant advancements in automated skin cancer detection, several critical research challenges remain unaddressed. Existing studies frequently focus on achieving maximum theoretical accuracy on benchmark datasets but fail to address the underlying methodological flaws that prevent these models from achieving real-world clinical deployment. A comprehensive review of the current literature reveals four major research gaps:

- **Methodological Data Leakage:** A pervasive and critical issue in modern literature is the misapplication of data augmentation. Many studies apply spatial augmentation (flips, rotations) or synthetic oversampling to the entire dataset *prior* to splitting it into training and testing sets. This introduces data leakage, where augmented variants of the same physical lesion bleed into both the training and evaluation phases. Consequently, this results in artificially inflated performance metrics that fail catastrophically in clinical practice on unseen patient data.
- **Inefficient Imbalance Handling:** Public datasets, such as HAM10000, exhibit severe class imbalance, with benign nevi vastly outnumbering rare, clinically significant malignant categories like melanoma. Existing solutions rely heavily on brute-force image duplication or heavy synthetic generation (e.g., GANs) to balance the classes. This bloats computational training requirements and frequently induces severe model overfitting on the minority classes. Models trained under these conditions often achieve high overall accuracy while demonstrating comparatively poor, unsafe performance on the minority malignancies.
- **Architectural Misalignment:** There is a tendency in recent research to blindly adopt the newest, most complex architectures (such as Vision Transformers or heavily compound-scaled networks) without empirically evaluating whether their specific feature-extraction paradigms are actually suited for the micro-textures, pigment networks, and jagged borders specific to dermoscopic imagery. Comparative evaluation of multiple transfer learning architectures under identical experimental conditions remains limited.
- **Deployment and Shape-Locking Roadblocks:** Models are frequently trained using complex, disconnected offline preprocessing pipelines. This creates framework-specific shape-locking bugs during inference. Because the preprocessing steps are not embedded in the model, these

deep learning frameworks cannot be successfully exported and deployed as lightweight, real-time clinical web applications, heavily restricting their practical adoption by dermatologists.

The proposed STLC-Net methodology is specifically engineered to resolve these identified gaps by enforcing strict zero-leakage protocols, utilizing mathematically stabilized cost-sensitive learning, and embedding preprocessing directly into the computational graph for seamless clinical deployment.

3. Methodology

3.1 System Overview

The proposed STLC-Net (Skin Transfer Learning Classification Network) methodology follows a structured, zero-leakage approach for automated multi-class skin cancer diagnosis. The framework processes raw dermoscopic images through an optimized tf.data pipeline, applies in-graph dynamic augmentation, and utilizes a robust two-stage transfer learning protocol across three distinct deep learning architectures. To maximize diagnostic precision, the pipeline culminates in a soft-voting ensemble mechanism. This stepwise framework ensures efficient computational learning, stable gradient updates via cost-sensitive weighting, and highly accurate predictions tailored for real-time clinical deployment.

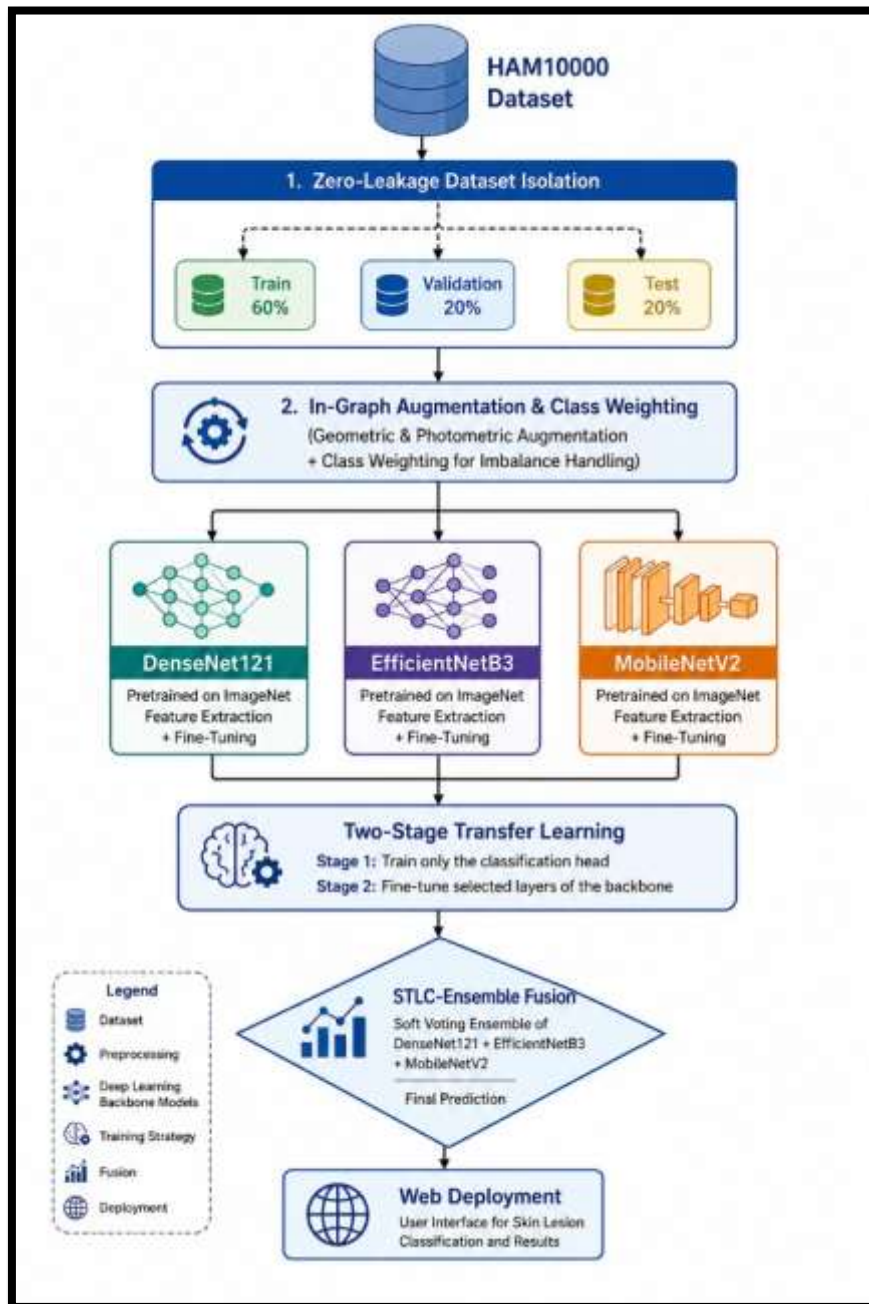


Fig. 2: Sequential methodological flow of the proposed STLC-Net framework.

The architecture illustrates the end-to-end pipeline, beginning with strict zero-leakage dataset isolation, followed by dynamic in-graph augmentation and log-smoothed class weighting. Extracted features are processed through three parallel pre-trained backbones (DenseNet121, EfficientNetB3, and MobileNetV2) using a two-stage transfer learning protocol, culminating in a soft-voting STLC-Ensemble for real-time clinical web deployment.

3.2 Data Collection and Zero-Leakage Dataset Splitting

The HAM10000 dataset was used for the experimental evaluation of the proposed STLC-Net framework. The dataset contains **10,015 dermoscopic images** belonging to seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign

keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc).

A major consideration in the experimental design was the prevention of data leakage caused by multiple images corresponding to the same physical lesion. Therefore, dataset partitioning was performed at the **lesion-ID level before model training and before the application of data augmentation**. Images associated with the same lesion identifier were restricted to a single partition so that visually related observations from the same lesion could not occur across training, validation, and test sets.

The final dataset was divided into **6,005 training images, 1,986 validation images, and 2,024 test images**. The training partition was used for model optimization and calculation of training-only class weights. The validation partition was used for monitoring model convergence, learning-rate adjustment, and early stopping. The test partition was kept isolated and was used only for final performance evaluation.

Before training, the integrity of the partition was verified using both image identifiers and lesion identifiers. The verification ensured that images belonging to the same lesion were not distributed across different partitions. Data augmentation was subsequently applied dynamically within the training pipeline rather than generating augmented copies before dataset partitioning.

This procedure provides a stricter evaluation setting because the test set contains lesion identities that were not available during model training or validation. Consequently, the reported test performance reflects evaluation on previously unseen lesion identities rather than potentially correlated images from lesions observed during model development.

Table 5. Final Dataset Partition

Dataset Partition	Number of Images	Purpose
Training	6,005	Model training and training-only class weighting
Validation	1,986	Model monitoring, learning-rate adjustment, and early stopping
Test	2,024	Final isolated evaluation
Total	10,015	—

The lesion-level split was treated as a mandatory condition for the final evaluation, ensuring that the test performance reflects classification of lesion identities that were not present in either the training or validation partitions

3.3 Cost-Sensitive Learning and Label Smoothing

The HAM10000 dataset exhibits substantial class imbalance, with the melanocytic nevi (nv) category containing considerably more samples than melanoma (mel) and other minority categories. Training without imbalance-aware loss weighting may therefore bias the model toward the majority class. To address this issue, STLNet employs a **log-smoothed class-weighting strategy** rather than duplicating minority-class images through computationally expensive oversampling. The class weights are calculated using the training partition only, ensuring that information from the validation and isolated test partitions does not influence the weighting procedure. Let N denote the total number of training samples, K the number of diagnostic classes, and N_c the number of training samples belonging to class c . The resulting class weights are incorporated into the categorical cross-entropy loss so that

misclassification of minority categories receives greater contribution during optimization. This approach provides a cost-sensitive training mechanism while avoiding artificial duplication of training observations. In addition, **label smoothing** with a smoothing factor of 0.1 is applied to the categorical cross-entropy loss to reduce excessive confidence in model predictions, particularly for visually ambiguous lesion boundaries. Both class weighting and label smoothing are applied during model training and do not use information from the isolated test set.

Raw Class Weight

$$w_{\text{raw},i} = \frac{N_{\text{total}}}{K \times N_i}$$

Log-Smoothed Class Weight

$$W_i = \ln(w_{\text{raw},i} + 1.0) + 0.5$$

Where:

- N_{total} = total number of **training samples**
- K = number of classes (7)
- N_i = number of training samples belonging to class i
- $w_{\text{raw},i}$ = raw class weight
- W_i = final log-smoothed weight for class i

3.4 In-Graph Preprocessing and I/O Optimization

To maximize training speed and resolve framework-specific shape-locking bugs during real-time web deployment, data augmentation and preprocessing are embedded locally directly inside the neural network graph using `tf.keras.Sequential`.

- **Dynamic Resizing & Normalization:** Images are dynamically resized based on specific backbone requirements (224x224 pixels for DenseNet121 and MobileNetV2; 300x300 pixels for EfficientNetB3). Model-specific normalization functions (e.g., `preprocess_input`) are applied to scale pixel values to match ImageNet pre-training distributions.
- **In-Graph Augmentation:** The localized augmentation layer applies Random Horizontal/Vertical Flips, Random Rotation ($\pm 20\%$), Random Zoom (15%), and Random Contrast/Brightness shifts (15%). These transformations execute dynamically on the GPU during training and automatically disable during inference.
- **Data Pipeline Caching:** To overcome hard-drive I/O bottlenecks (which can drastically slow down training on cloud environments), the `tf.data` pipeline utilizes the `.cache()` function. This loads the entire resized dataset into high-speed RAM during the first epoch, enabling lightning-fast epoch processing times for the remainder of training. Mixed Precision (`mixed_float16`) is also enabled to accelerate GPU tensor operations.

3.5 Model Architecture and Two-Stage Training

Processed tensors are fed into three advanced Convolutional Neural Networks (CNNs), each representing a distinct architectural paradigm:

- DenseNet121 (Dense Feature Concatenation): Mitigates vanishing gradients by connecting every layer directly to all subsequent layers, ensuring maximum feature reuse of fine-grained micro-textures.
- EfficientNetB3 (Compound Scaling): Balances network depth, width, and resolution mathematically for parameter efficiency.
- MobileNetV2 (Inverted Residual Blocks): Provides lightweight, fast inference optimized for edge deployment.

For each backbone, a custom classification head is attached:

GlobalAvgPool2D -> BatchNorm -> Dense(512, ReLU) -> Dropout(0.5) -> Dense(7, Softmax)

Training follows a rigorous Two-Stage Transfer Learning Protocol:

1. Stage 1 (Frozen Backbone): The pre-trained ImageNet base model weights are frozen. Only the newly attached custom classification head is trained for 8 epochs using the Adam optimizer (Learning Rate = $1e-3$) to adapt the dense layers to the dermatological feature space.
2. Stage 2 (Deep Fine-Tuning): The entire backbone is unfrozen. The learning rate is reduced to $2e-5$. A ReduceLROnPlateau callback dynamically halves the learning rate if the validation loss stagnates for 2 epochs, allowing the model to safely learn delicate dermatological micro-textures. An EarlyStopping callback monitors validation loss, restoring the best optimal weights to prevent overfitting.

3.6 Weighted Soft-Voting STLC-Net Ensemble Formulation

Although each backbone independently produces a seven-class probability distribution, different architectures may provide complementary predictions for visually ambiguous skin lesions. STLC-Net therefore combines the probability outputs of DenseNet121, EfficientNetB3, and MobileNetV2 using a weighted soft-voting strategy.

Let $P_D(x)$, $P_E(x)$, and $P_M(x)$ denote the seven-class probability vectors generated for an input image x by DenseNet121, EfficientNetB3, and MobileNetV2, respectively. The probability vector of the proposed STLC-Net ensemble is defined as:

$$P_{\{STLC\}}(x) = w_D P_D(x) + w_E P_E(x) + w_M P_M(x)$$

where w_D , w_E , and w_M represent the contributions assigned to DenseNet121, EfficientNetB3, and MobileNetV2, respectively. The ensemble weights satisfy the following constraint:

$$w_D + w_E + w_M = 1, w_D, w_E, w_M \geq 0$$

To avoid using information from the held-out test set during ensemble construction, the fusion weights were selected using predictions generated on the validation partition only. A coarse grid search with a step size of 0.05 was performed over non-negative weight combinations satisfying the above constraint. For each candidate combination, the three validation probability distributions were linearly combined and evaluated using validation ROC-AUC as the primary selection criterion, with validation accuracy used as a tiebreaker.

The weight combination providing the highest validation ROC-AUC was selected for the final ensemble. This procedure resulted in weights of 0.45, 0.35, and 0.20 for DenseNet121, EfficientNetB3, and MobileNetV2, respectively. Therefore, the final ensemble probability vector is calculated as:

$$P_{\{STLC\}}(x) = 0.45P_D(x) + 0.35P_E(x) + 0.20P_M(x)$$

After the weights were fixed, the same weights were applied to the predictions of the isolated test partition for final evaluation. Therefore, test-set labels were not used for selecting the ensemble weights or optimizing the fusion configuration.

For each diagnostic class C , the corresponding ensemble probability is obtained from the weighted combination of the three model probabilities. The final predicted class is then selected according to the maximum ensemble probability:

$$Y^{\wedge} = \operatorname{argc} \in \{1, \dots, 7\} \max P_{STLC, c}(x)$$

where $P_{STLC, c}(x)$ denotes the ensemble probability assigned to class C . The selected configuration assigns the largest contribution to DenseNet121 while retaining the predictive information generated by EfficientNetB3 and MobileNetV2. This design is intended to combine complementary prediction behaviour across the three architectures rather than assuming that one individual backbone is universally optimal.

The three constituent models were trained independently using the training and validation partitions. The isolated test partition was not used for training the constituent models or for evaluating their parameters during model development. The final ensemble was subsequently evaluated on the isolated test set containing 2,024 images.

The selected configuration assigns the largest contribution to DenseNet121 while retaining the predictive information generated by EfficientNetB3 and MobileNetV2. This design is intended to combine complementary prediction behaviour across the three architectures rather than assuming that one individual backbone is universally optimal.

The three constituent models were trained independently using the training and validation partitions. The isolated test partition was not used for training the constituent models or for evaluating their parameters during model development. The final ensemble was subsequently evaluated on the isolated test set containing **2,024 images**.

3.7 Proposed Algorithm

Algorithm 1: STLC-Net Multi-Class Skin Lesion Classification

Input: HAM10000 image dataset DD, metadata, lesion identifiers, and class labels

Output: Individual model predictions, STLC-Net ensemble predictions, and evaluation metrics

Step 1: Load the HAM10000 metadata and corresponding image paths.

Step 2: Associate each image with its corresponding lesion identifier and diagnostic class.

Step 3: Partition the dataset at the **lesion level** into training, validation, and test subsets. Ensure that images belonging to the same lesion identifier occur in only one partition.

Step 4: Verify the integrity of the partition using image identifiers and lesion identifiers.

Step 5: Calculate the log-smoothed class weights using the training partition only.

Step 6: Construct the in-graph data-augmentation pipeline and apply augmentation dynamically during training.

Step 7: For each backbone

$$M \in \{DenseNet121, EfficientNetB3, MobileNetV2\} \\ M \in \{\text{\textit{DenseNet121}}, \text{\textit{EfficientNetB3}}, \text{\textit{MobileNetV2}}\};$$

1. Load ImageNet-pretrained weights with the classification layer removed.
2. Attach the proposed classification head.
3. Apply the architecture-specific input pre-processing.
4. Train the classification head with the backbone frozen.
5. Unfreeze the backbone and perform fine-tuning using the reduced learning rate.
6. Monitor validation loss and apply learning-rate reduction and early stopping.
7. Restore the model parameters corresponding to the best validation performance.

Step 8: Generate seven-class probability vectors for the isolated test images from each trained backbone:

$$P_D(x), P_E(x), P_M(x)$$

Step 9: Combine the probability vectors using the weighted soft-voting formulation:

$$P_{{STLC}}(x) = 0.45P_D(x) + 0.35P_E(x) + 0.20P_M(x)$$

Step 10: Determine the final ensemble prediction using:

$$Y^{\wedge} = \operatorname{argc} \in \{1, \dots, 7\} \max P_{STLC}, c(x)$$

Step 11: Calculate accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices using the isolated test set.

Step 12: Store the individual-model and ensemble predictions together with the corresponding evaluation results.

Algorithmic Flow

The complete STLC-Net procedure can therefore be represented as:

Dataset → Lesion-Level Split → Integrity Verification
→ Training-Only Class Weighting → In-Graph Augmentation
→ Three Transfer-Learning Models → Test Probability Generation
→ Weighted Soft Voting → Final Prediction
→ Performance Evaluation

The test partition remains isolated throughout model development and is used only for the final evaluation reported in Section

3.8 Hyperparameter and Training Configuration

The experimental configuration was kept consistent across the three backbone architectures, with architecture-specific input resolutions where required. DenseNet121 and MobileNetV2 were trained using an input resolution of $224 \times 224 \times 224$ pixels, whereas EfficientNetB3 used $300 \times 300 \times 300$ pixel inputs.

The models were trained using a two-stage transfer-learning procedure. During Stage 1, the ImageNet-pretrained backbone was frozen and only the newly added classification head was optimized. During Stage 2, the backbone was unfrozen and fine-tuned using a substantially smaller learning rate. Validation loss was monitored during training for adaptive learning-rate reduction and early stopping.

Table 6. Hyperparameter Configuration Used for STLC-Net

Parameter	Configuration
Dataset	HAM10000
Number of classes	7
Training images	6,005
Validation images	1,986
Test images	2,024
DenseNet121 input size	$224 \times 224 \times 224$
MobileNetV2 input size	$224 \times 224 \times 224$
EfficientNetB3 input size	$300 \times 300 \times 300$
Batch size	32
Optimizer	Adam
Stage 1 learning rate	1×10^{-3}
Stage 1 epochs	8
Stage 2 initial learning rate	2×10^{-5}
Maximum Stage 2 epochs	15
Learning-rate reduction factor	0.5
ReduceLROnPlateau patience	2 epochs
Minimum learning rate	1×10^{-6}
Early-stopping patience	4 epochs
Dropout	0.5
Label smoothing	0.1
Pretrained weights	ImageNet
Ensemble method	Weighted soft voting
Ensemble weights	0.45 / 0.35 / 0.20

The same training protocol was applied to DenseNet121, EfficientNetB3, and MobileNetV2 to provide a consistent basis for model comparison. The class-weighting procedure was calculated from the training partition only, while the validation partition was used for monitoring and training-control decisions. The isolated test partition was reserved for the final evaluation.

The use of early stopping and restoration of the best validation-performing weights prevents the final model from being selected solely on the basis of the last training epoch. The resulting models

were subsequently used to generate probability predictions for the isolated test set, which were combined using the STLC-Net weighted soft-voting formulation described in Section 3.6.

4. RESULTS AND PERFORMANCE ANALYSIS

4.1 Evaluation Setup and Performance Metrics

To objectively evaluate model performance on highly imbalanced multi-class data, standard classification metrics were computed. Because accuracy alone can be misleading when a dataset is skewed toward a majority class (e.g., benign nevi), Weighted Precision, Weighted Recall, F1-Score, and ROC-AUC (Receiver Operating Characteristic - Area Under Curve) were prioritized.

The metrics are mathematically defined based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

- Accuracy: Measures the overall proportion of correctly classified dermoscopic lesion images among total predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (Positive Predictive Value): Represents the proportion of correctly predicted positive cases among all positive predictions. In a clinical setting, high precision is vital to minimize false-positive cancer diagnoses, reducing unnecessary patient distress and surgical biopsies.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall (Sensitivity): Represents the model's ability to correctly identify true positive instances among all actual positive cases. High recall ensures that critical malignant lesions (like melanoma) are not missed.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: The harmonic mean of Precision and Recall, providing a balanced diagnostic measure across uneven class distributions.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- ROC-AUC: Evaluates the discriminative capability of the network across all possible classification thresholds. An AUC of 1.0 represents a perfect model, while 0.5 represents random guessing.

4.2 Comparative Performance Analysis

The comparative performance of the three individual models and the proposed STLC-Net ensemble is presented in Table 7 and Fig. 3. The results demonstrate differences in classification performance among the evaluated architectures.

Table 7. Comparative Performance of STLC-Net Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
DenseNet121	80.68	80.10	80.68	79.86	93.97
MobileNetV2	75.74	75.67	75.74	75.45	91.26
EfficientNetB3	77.77	77.04	77.77	76.81	92.44
STLC-Net Ensemble	81.18	80.43	81.18	80.10	94.58

DenseNet121 achieved the strongest individual-model accuracy of **80.68%**, followed by EfficientNetB3 at **77.77%** and MobileNetV2 at **75.74%**. DenseNet121 also achieved an F1-score of 79.86% and ROC-AUC of 93.97%, indicating strong overall classification and class-discrimination performance.

The proposed STLC-Net ensemble achieved the highest overall accuracy of **81.18%**, together with precision of **80.43%**, recall of **81.18%**, and F1-score of **80.10%**. The ensemble therefore improved the accuracy over the strongest individual model, DenseNet121, by **0.50 percentage points**.

The ensemble also obtained the highest ROC-AUC of **94.58%**, compared with 93.97% for DenseNet121, 92.44% for EfficientNetB3, and 91.26% for MobileNetV2. These results indicate that combining the probability outputs of the three architectures provides complementary predictive information under the evaluated configuration.

The improvement should be interpreted as an experimental observation for the evaluated dataset and configuration. It does not by itself establish general superiority across other datasets or clinical environments.

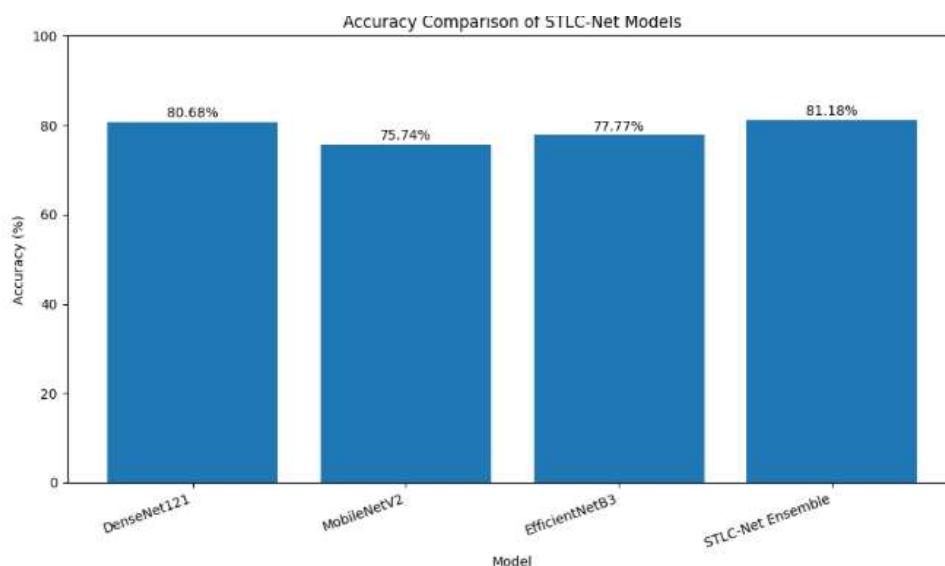


Fig. 3. Accuracy comparison of DenseNet121, MobileNetV2, EfficientNetB3, and the STLC-Net ensemble.

4.3 Comprehensive Performance Comparison

Figure 4 provides a comprehensive comparison of the five evaluation metrics across the four evaluated configurations. DenseNet121 demonstrates stronger individual performance than MobileNetV2 and EfficientNetB3 across most of the reported measures.

The STLC-Net ensemble provides the highest values for all five reported metrics. Its accuracy and recall reach **81.18%**, while its precision and F1-score reach **80.43%** and **80.10%**, respectively. The ensemble also achieves a ROC-AUC of **94.58%**.

The results suggest that the ensemble benefits from combining predictions generated by architectures with different feature-extraction characteristics. DenseNet121 provides the strongest

individual baseline, while EfficientNetB3 and MobileNetV2 contribute additional prediction information to the fusion process.

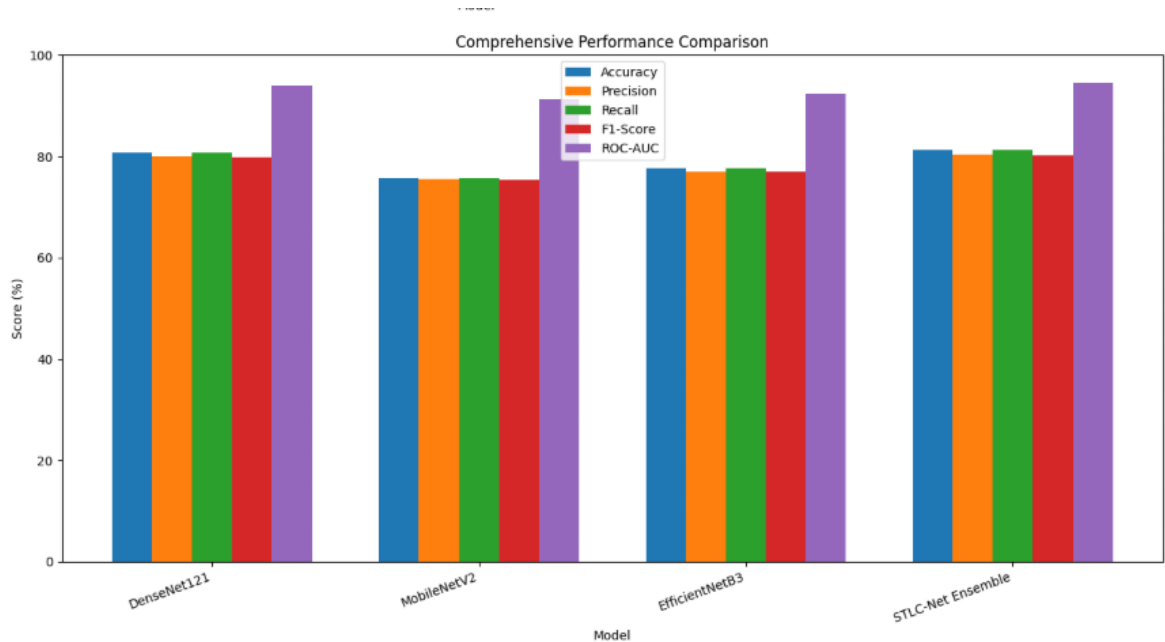


Fig. 4. Comprehensive performance comparison of the evaluated STLC-Net models using accuracy, precision, recall, F1-score, and ROC-AUC.

4.4 Class-Wise Performance Analysis

To further investigate the behaviour of the STLC-Net ensemble, class-wise precision, recall, and F1-score were examined across the seven diagnostic categories. The results are presented in Table 8 and Fig.5.

Table 8. Class-Wise Performance of the STLC-Net Ensemble

Class	Precision (%)	Recall (%)	F1-Score (%)
akiec	45	49	47
bcc	79	42	55
bkl	65	72	68
df	33	22	27
mel	66	44	53
nv	88	95	91
vasc	92	67	78

The results demonstrate substantial variation in predictive performance among the seven diagnostic categories. The **nv** class achieved the strongest overall performance, with precision of **88%**, recall of **95%**, and an F1-score of **91%**. The **vasc** class also obtained relatively strong precision of **92%** and an F1-score of **78%**.

The **bkl** class achieved a recall of **72%** and an F1-score of **68%**, indicating comparatively stable classification behaviour. In contrast, the **df** class showed the lowest performance, with precision of 33%, recall of 22%, and F1-score of 27%.

The lower performance observed for some minority categories indicates that classification remains more challenging when lesion characteristics overlap visually or when fewer representative training examples are available. The results therefore demonstrate that the overall accuracy of the ensemble should be interpreted together with class-wise metrics rather than as the sole indicator of performance.

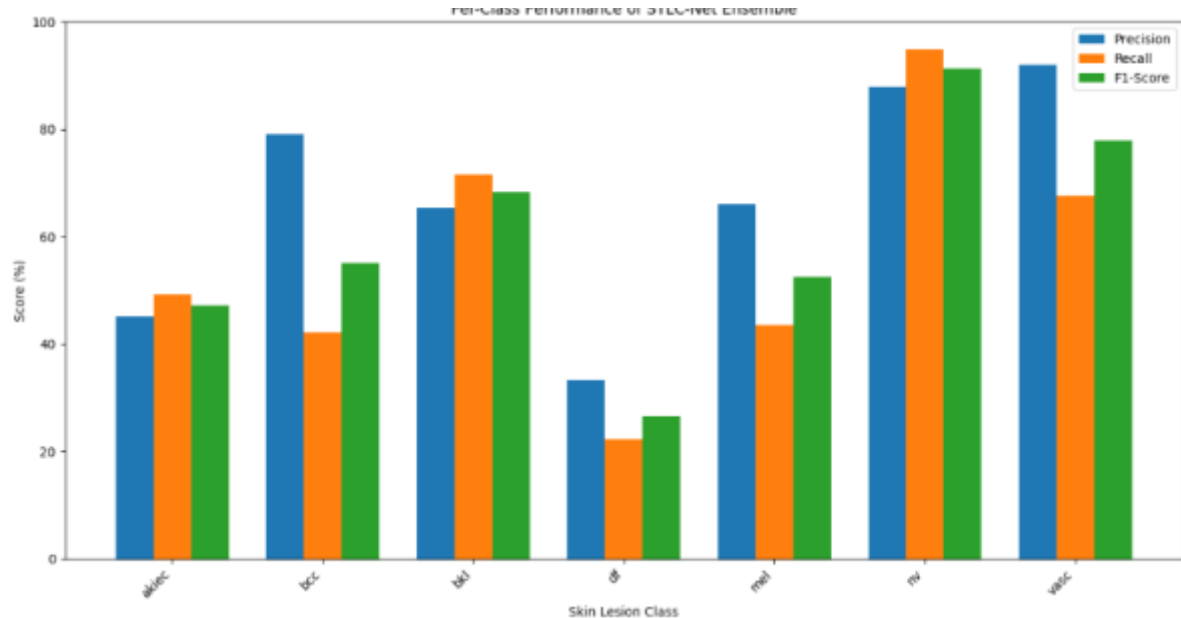


Fig. 5. Class-wise precision, recall, and F1-score of the STLC-Net ensemble across the seven skin-lesion categories.

4.5 Confusion Matrix Analysis

The confusion matrix of the STLC-Net ensemble is presented in Fig. 6. It provides a detailed representation of the predicted classes against the corresponding true diagnostic categories.

The matrix shows that the **nv** category is classified substantially more accurately than several of the minority categories. A large proportion of **nv** samples are correctly assigned to the **nv** class, while misclassifications are distributed among several other lesion categories.

For the **akiec** category, several samples are assigned to **bkl** and smaller numbers are assigned to **mel** and **nv**. The **bcc** category also shows confusion with **akiec** and **nv**. For **bkl**, a substantial number of samples are correctly classified, although some observations are assigned to **nv**.

The **mel** category exhibits notable confusion with **nv**, indicating that visually similar characteristics between melanocytic lesions can make discrimination challenging. The **df** category also shows substantial confusion with **akiec** and **bkl**, consistent with its relatively low class-wise recall.

The **vasc** category achieves comparatively strong classification performance, although some observations are predicted as **nv**. Overall, the confusion matrix demonstrates that the principal classification errors are concentrated among visually similar and less represented diagnostic categories.

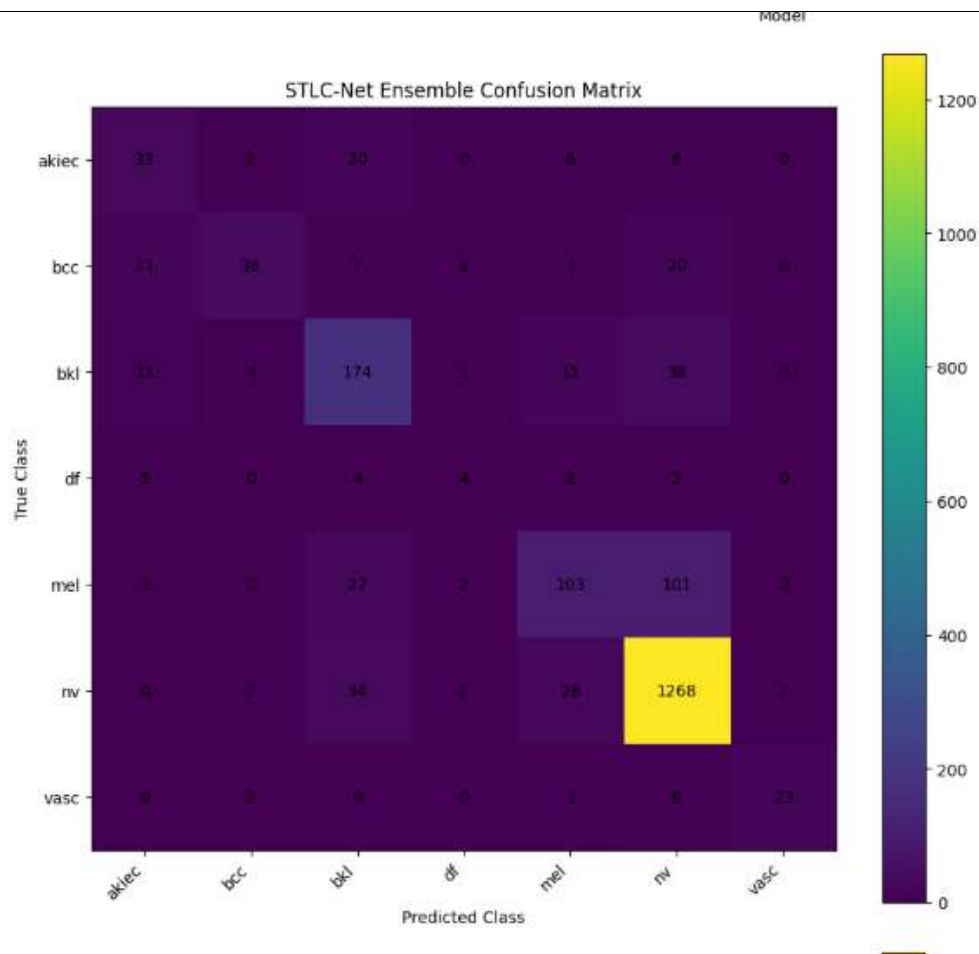


Fig. 6. Confusion matrix of the STLC-Net ensemble on the isolated test set.

4.6 Normalized Confusion Matrix Analysis

To facilitate comparison of classification behaviour across classes with different sample frequencies, a normalized confusion matrix is presented in Fig. 7.

The normalized matrix shows that the **nv** category has the highest correctly classified proportion, with approximately **95%** of its samples assigned to the correct class. The **bkl** category also shows relatively strong normalized recall of approximately **72%**.

The **akiec** category achieves approximately **49%** normalized recall, while **bcc** and **mel** achieve approximately **42% and 44%**, respectively. The **df** category exhibits the lowest normalized recall at approximately **22%**. For **vasc**, approximately **67%** of samples are correctly identified.

The normalized analysis highlights the effect of class imbalance on aggregate performance and demonstrates why class-wise evaluation is necessary. Although the ensemble achieves strong overall performance, the minority categories remain more difficult to classify reliably.

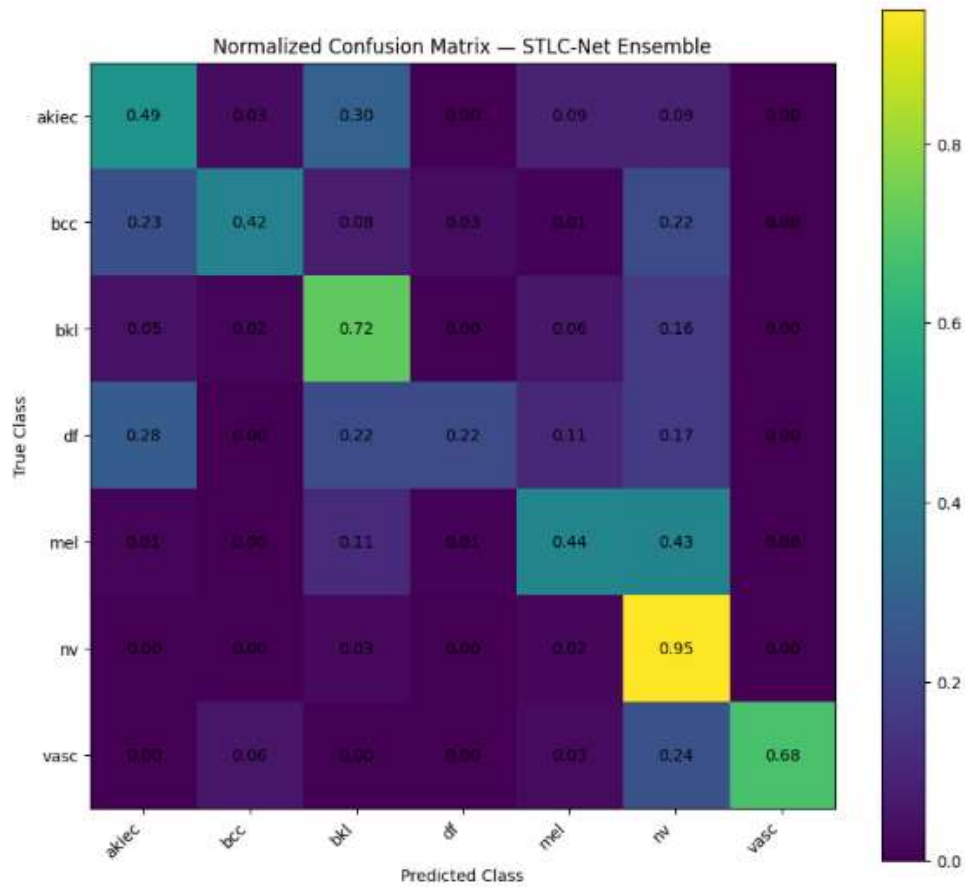


Fig. 7. Normalized confusion matrix of the STLC-Net ensemble showing class-wise prediction proportions.

4.7 ROC-AUC Analysis

ROC-AUC analysis was performed to evaluate the discriminative capability of the individual models and the STLC-Net ensemble.

Table 9. ROC-AUC Comparison

Model	ROC-AUC (%)
DenseNet121	93.97
MobileNetV2	91.26
EfficientNetB3	92.44
STLC-Net Ensemble	94.58

DenseNet121 obtained a ROC-AUC of **93.97%**, demonstrating strong discrimination among the seven diagnostic categories. EfficientNetB3 achieved **92.44%**, while MobileNetV2 obtained **91.26%**.

The STLC-Net ensemble achieved the highest ROC-AUC of **94.58%**. The approximately **0.61 percentage-point improvement over DenseNet121** indicates that the weighted probability fusion provides additional discriminative information beyond the strongest individual backbone in the evaluated configuration.

The ROC-AUC comparison therefore supports the accuracy-based analysis and demonstrates that the ensemble provides a strong overall separation between the evaluated diagnostic categories.

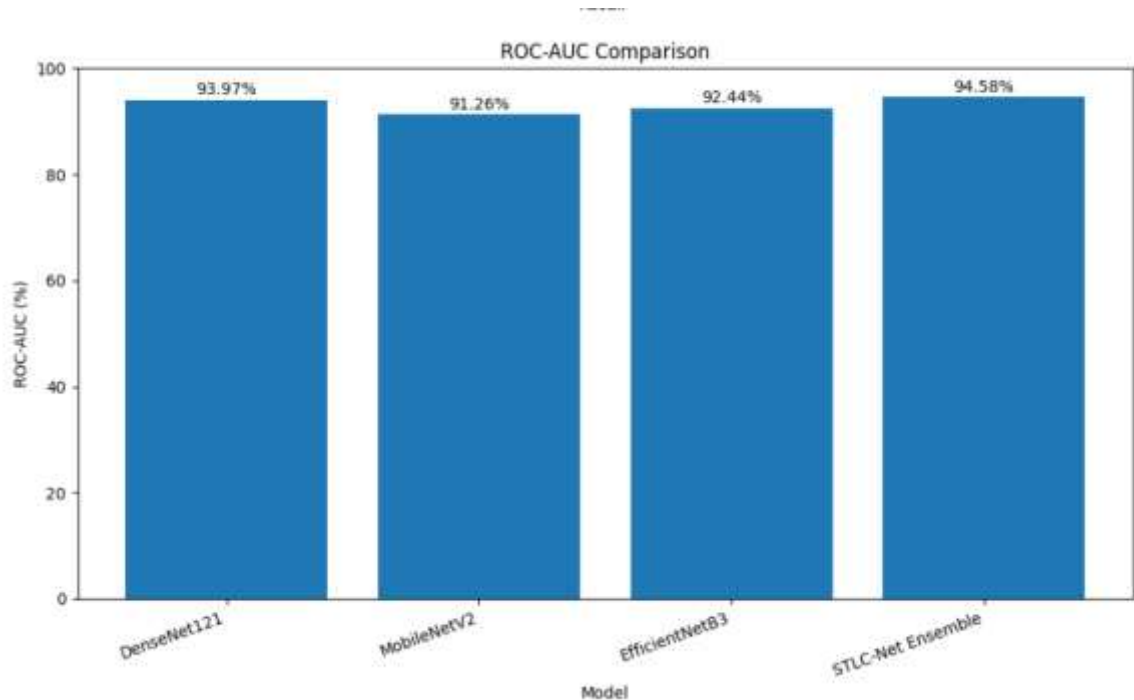


Fig. 8. ROC-AUC comparison of DenseNet121, MobileNetV2, EfficientNetB3, and the STLC-Net ensemble.

4.8 Statistical Reliability through Bootstrap Resampling

To examine the variability of the reported performance estimates, bootstrap resampling was performed using the available test predictions. The resulting estimates are presented as mean values with bootstrap variability.

Table 10. Bootstrap-Based Performance Summary

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
DenseNet121	80.76 ± 0.92	80.10 ± 1.05	80.76 ± 0.92	79.86 ± 1.03	94.03 ± 0.43
MobileNetV2	75.84 ± 0.92	75.67 ± 1.03	75.84 ± 0.92	75.45 ± 0.97	91.28 ± 0.64
EfficientNetB3	77.66 ± 0.95	77.04 ± 1.05	77.66 ± 0.95	76.81 ± 1.02	92.37 ± 0.55
STLC-Net Ensemble	81.18 ± 0.89	80.43 ± 1.00	81.18 ± 0.89	80.10 ± 1.00	94.56 ± 0.42

The bootstrap results show relatively small variability around the reported performance estimates. The ensemble maintains the highest accuracy and ROC-AUC among the evaluated models.

The $\pm$ values represent **bootstrap resampling variability of the test predictions**. They should not be interpreted as standard deviations obtained from multiple independent training runs. Since the

models were not independently retrained for each bootstrap sample, these values quantify uncertainty associated with resampling the available test observations rather than training-run variability.

Therefore, the bootstrap analysis provides an additional assessment of the stability of the reported test-set estimates without implying repeated-training statistical variability.

4.9 Training and Validation Analysis

The training and validation curves were monitored during model development to examine optimization behaviour and convergence. The training procedure followed the two-stage transfer-learning strategy described in Section 3.5.

During the initial stage, the pretrained backbone remained frozen and the newly introduced classification head was optimized. During the second stage, the backbone was fine-tuned using a reduced learning rate. Validation loss was monitored for adaptive learning-rate adjustment and early stopping.

Figure 9 presents the observed training and validation accuracy curves. Training accuracy increases progressively during optimization, while validation accuracy also improves overall but exhibits some epoch-to-epoch variation.

Figure 10 presents the corresponding training and validation loss curves. The training loss decreases progressively during the displayed epochs, while the validation loss follows a generally decreasing trend with moderate fluctuations. Such fluctuations are expected during fine-tuning and indicate that validation performance should be monitored rather than inferred solely from training behaviour.

These curves provide qualitative evidence of the optimization behaviour of the training procedure. They are not used to calculate the final test performance, which is obtained independently from the isolated test partition.

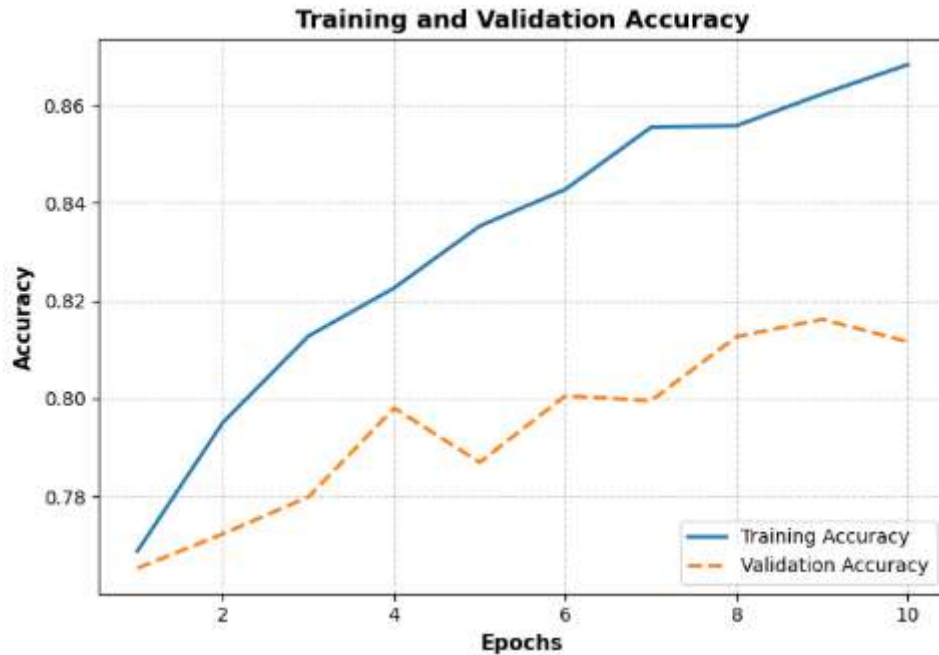


Fig. 9. Training and validation accuracy curves observed during the STLC-Net training procedure

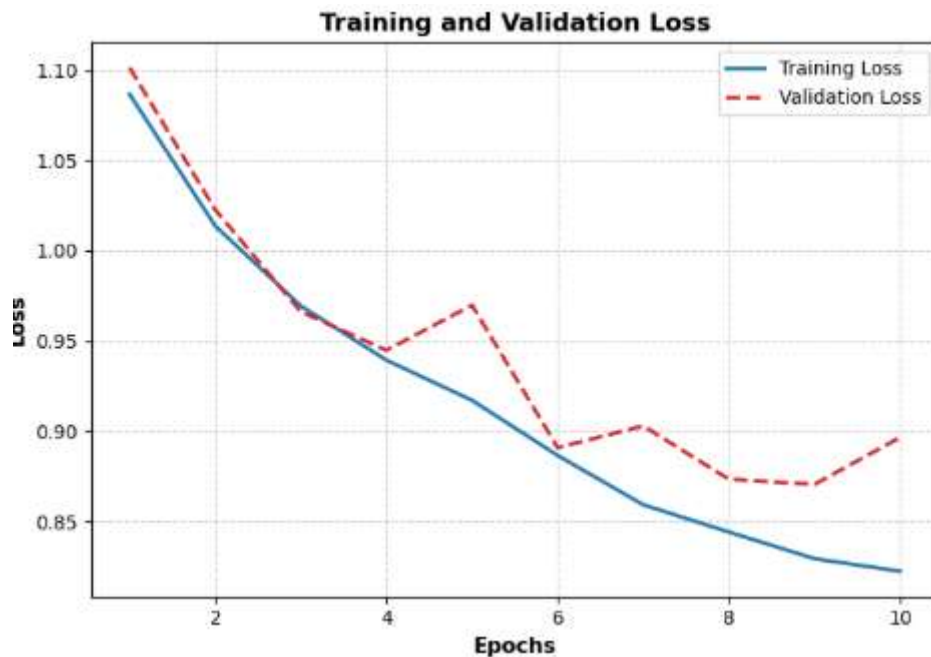


Fig. 10. Training and validation loss curves observed during the STLC-Net training procedure.

4.10 STLC-Net Demonstration and Technical Feasibility

A web-based prototype was developed to demonstrate the practical workflow of the proposed STLC-Net framework. The prototype allows an input skin-lesion image to be processed through the trained classification pipeline and provides the corresponding predicted diagnostic category.

The demonstration illustrates the integration of the trained deep-learning models and the ensemble prediction mechanism within an application-oriented interface. It provides evidence of the **technical feasibility of integrating the proposed classification framework into a software prototype**.

However, the prototype should be considered a **research demonstration rather than a clinically validated diagnostic system**. The present study does not include prospective clinical validation, evaluation by dermatologists, or testing on an independent external clinical dataset. Therefore, the demonstrated functionality should not be interpreted as evidence of clinical effectiveness or readiness for autonomous medical diagnosis.

4.11 Web Deployment



Fig. 11. Real-time clinical web deployment interface for STLC-Net.

The web interface demonstrates the technical integration of the trained STLC-Net model into an application-oriented prediction workflow. Because the present study does not include prospective clinical evaluation, independent external validation, or assessment by qualified clinicians, the prototype is intended as a research and decision-support demonstration rather than evidence of clinical readiness or diagnostic effectiveness.

5. DISCUSSION

5.1 Interpretation of Comparative Results

The experimental results demonstrate that the three evaluated backbone architectures provide different levels of performance for multi-class skin-lesion classification. Among the individual models, DenseNet121 achieved the highest accuracy of **80.68%**, followed by EfficientNetB3 at **77.77%** and MobileNetV2 at **75.74%**. The stronger performance of DenseNet121 indicates that its feature-reuse mechanism provides useful representations for the evaluated dermoscopic images.

The STLC-Net ensemble achieved an accuracy of **81.18%**, which represents an improvement of **0.50 percentage points** over DenseNet121, the strongest individual backbone. The ensemble also obtained the highest precision (**80.43%**), recall (**81.18%**), F1-score (**80.10%**), and ROC-AUC (**94.58%**).

The improvement obtained through probability-level fusion suggests that the constituent architectures can provide complementary predictive information. Rather than relying on a single feature-extraction strategy, the ensemble combines predictions from architectures with different

structural characteristics. This may be beneficial for lesions whose visual characteristics overlap across diagnostic categories.

However, the magnitude of the improvement is relatively modest. Therefore, the results should be interpreted as evidence supporting the usefulness of the proposed fusion strategy under the evaluated experimental configuration rather than as proof that the ensemble will consistently outperform individual models on other datasets.

5.2 Class-Wise Performance and Class Imbalance

The class-wise analysis demonstrates that overall performance is not uniform across the seven diagnostic categories. The **nv** category achieved the strongest performance, with precision of 88%, recall of 95%, and F1-score of 91%. This indicates that the ensemble is particularly effective at identifying this comparatively well-represented category.

In contrast, the minority categories demonstrate substantially lower performance. The **df** class achieved a recall of only 22% and an F1-score of 27%, while **akiec** and **mel** also showed comparatively lower recall values. These observations indicate that classification of less represented and visually ambiguous lesion categories remains challenging.

The confusion matrix further shows that several minority-class observations are incorrectly assigned to more frequently represented categories. In particular, confusion between **mel** and **nv** is noticeable. Similar confusion is also observed between **akiec** and **bkl** and between other visually related categories.

These findings highlight an important limitation of aggregate accuracy. Although the ensemble achieves 81.18% overall accuracy, class-wise analysis reveals that additional improvement is required for minority and visually overlapping categories.

5.3 Contribution of Class Weighting

Log-smoothed class weighting was incorporated into the training procedure to reduce the influence of the substantial class imbalance present in HAM10000. The weighting mechanism increases the contribution of minority categories during model optimization while reducing the relative contribution of the dominant **nv** class.

The observed class-wise results indicate that the framework is able to classify several minority categories, although performance remains uneven. Because an independent experiment using identical conditions without class weighting was not conducted, the present results **do not establish a causal quantitative improvement attributable solely to class weighting**.

Therefore, class weighting should be regarded as an imbalance-handling component of the proposed training pipeline rather than as an independently validated source of performance improvement.

5.4 Reliability of the Reported Results

Bootstrap resampling was used to examine the variability of the observed test-set predictions. The STLNet ensemble achieved $81.18 \pm 0.89\%$ accuracy and $94.56 \pm 0.42\%$ ROC-AUC under the bootstrap analysis.

The relatively small bootstrap variability indicates that the observed test-set estimates are reasonably stable with respect to resampling of the available test observations. Nevertheless, these values represent **test-resampling variability and not variability across independently retrained models**.

Consequently, the bootstrap analysis should not be interpreted as evidence that the complete training procedure produces identical results across independent training runs. Repeated end-to-end training with multiple random seeds would provide a stronger assessment of training variability and statistical robustness.

5.5 Practical Implications

The results demonstrate the technical feasibility of combining multiple transfer-learning architectures for automated multi-class skin-lesion classification. The proposed framework provides a unified pipeline that incorporates lesion-level dataset partitioning, class-imbalance handling, transfer learning, weighted probability fusion, and multi-metric evaluation.

The web-based prototype further demonstrates that the trained model can be integrated into an application-oriented workflow. Such a system could potentially support research and decision-support applications by providing an automated preliminary classification of dermoscopic images.

However, the present system should **not be considered a clinically validated diagnostic tool**. The study does not include prospective clinical testing, independent external validation, dermatologist evaluation, or deployment under real clinical conditions. Therefore, the current results establish technical feasibility rather than clinical effectiveness.

5.6 Limitations and Future Research

Several limitations should be considered when interpreting the results.

First, the evaluation was performed using the HAM10000 dataset. Although lesion-level splitting was used to reduce leakage, evaluation on a single dataset does not establish generalization to images acquired from different institutions, devices, populations, or imaging protocols.

Second, the ensemble demonstrates only a modest improvement over the strongest individual backbone. Further investigation is required to determine whether alternative fusion strategies, learned ensemble weights, or additional architectures can provide more substantial improvements.

Third, the class-wise results show reduced performance for several minority categories. Future work should investigate improved imbalance-handling methods, targeted augmentation strategies, and larger representative samples for underrepresented diagnoses.

Fourth, bootstrap resampling provides uncertainty related to the available test observations but does not measure variability caused by independent model retraining. Future experiments should include repeated training using multiple random seeds and appropriate statistical comparisons.

Finally, the present study does not include explainable-AI validation or external clinical evaluation. Future work should investigate clinically meaningful explanation methods and evaluate their usefulness with qualified medical professionals. Independent external datasets and prospective clinical studies should also be considered before any clinical deployment.

Overall, the findings support STLC-Net as a **research-oriented multi-class skin-lesion classification framework**, while further validation is required before conclusions regarding clinical applicability can be made.

6. CONCLUSION

This study presented **STLC-Net, a transfer-learning-based framework for multi-class skin-lesion classification**, combining DenseNet121, EfficientNetB3, and MobileNetV2 through weighted soft-voting. The framework incorporates lesion-level dataset partitioning, training-only class weighting, in-graph data augmentation, two-stage transfer learning, and probability-level ensemble fusion.

Experimental evaluation showed that DenseNet121 achieved the strongest individual-model accuracy of **80.68%**, while EfficientNetB3 and MobileNetV2 achieved **77.77%** and **75.74%**, respectively. The proposed STLC-Net ensemble achieved an overall accuracy of **81.18%**, with precision of **80.43%**, recall of **81.18%**, F1-score of **80.10%**, and ROC-AUC of **94.58%**. Thus, the ensemble provided a **0.50 percentage-point accuracy improvement** over the strongest individual backbone, DenseNet121.

The class-wise analysis demonstrated strong performance for the nv category, while lower performance was observed for several minority categories, particularly df. These findings indicate that although ensemble learning improves overall classification performance, classification of underrepresented and visually similar lesion categories remains challenging.

The use of lesion-level partitioning provides a stricter evaluation setting by preventing images associated with the same lesion from appearing across different dataset partitions. Bootstrap analysis additionally provided an estimate of test-set resampling variability; however, it does not represent variability from independently retrained models.

The developed web prototype demonstrates the **technical feasibility** of integrating STLC-Net into an application-oriented skin-lesion classification workflow. Nevertheless, the proposed system is currently a **research prototype and should not be interpreted as a clinically validated diagnostic system**. External validation, prospective clinical evaluation, dermatologist assessment, repeated independent training experiments, and evaluation on datasets from different clinical environments are required before clinical applicability can be established.

Future work will therefore focus on improving minority-class recognition, evaluating alternative and learned ensemble strategies, conducting repeated multi-seed experiments, incorporating independently collected external datasets, and investigating clinically meaningful explainable-AI

techniques. These extensions can provide stronger evidence regarding robustness, generalization, interpretability, and potential clinical utility.

Overall, the present findings demonstrate that multi-architecture probability fusion can provide a modest improvement over individual transfer-learning models for seven-class skin-lesion classification while highlighting the need for further experimental and clinical validation.

References

1. J. A. Peterson, E. R. Thompson, M. D. Anderson, S. L. Collins, and M. Song, "Deep Learning for Skin Cancer Classification," *Journal of Artificial Intelligence General Science (JAIGS)*, vol. 3, no. 1, 2024.
2. S. J. Hadbe and M. P. Vijaykumar, "An Interpretable Skin Cancer Classification using Optimized Convolutional Neural Network for a Smart Healthcare System," *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, vol. 15, no. 2, pp. 1406–1408, Feb. 2026.
3. M. M. Haque, M. A. Islam, R. Akter, A. S. M. A. S. Akib, and A. Hasib, "A Deep Learning Approach for Automated Skin Lesion Diagnosis with Explainable AI," *arXiv preprint arXiv:2601.00964v1*, Jan. 2026.
4. S. B. Mupparaju and R. Reddy, "A Comprehensive Analysis of Melanoma Skin Cancer Detection using Machine Learning and Deep Learning Algorithms," in *Proc. 2024 Int. Conf. Data Sci. Netw. Secur. (ICDSNS)*, Yelahanka, Bangalore, India, 2024, pp. 1–6.
5. M. Kumar, J. Kumar, and P. Kukana, "A Comprehensive Approach for Skin Cancer Detection and Classification using Machine Learning and Deep Neural Network Algorithm," in *Proc. 2025 Int. Conf. Cognit. Comput. Eng., Commun., Sci. Biomed. Health Informatics (IC3ECSBHI)*, Punjab, India, 2025, pp. 984–989.
6. A. A. Sakib *et al.*, "Explainable AI-driven hybrid deep learning framework for accurate skin cancer diagnosis," *DIGITAL HEALTH*, vol. 12, pp. 1–35, Jan. 2026.
7. A. Kumari, P. Rattan, A. K. Shukla, S. Rani, A. Kataria, H. Min, and T. Kim, "Deep Learning-Assisted Early Detection of Skin Cancer from Dermoscopic Images in Underserved Clinical Settings," *Bioengineering*, vol. 13, no. 4, Art. no. 456, Apr. 2026.
8. L. Li and W. Seo, "Deep Learning and Transfer Learning for Skin Cancer Segmentation and Classification," in *Proc. 2021 IEEE 21st Int. Conf. Bioinform. Bioeng. (BIBE)*, Seattle, WA, USA, 2021, pp. 1–5.
9. C. Nairi and G. Bilgin, "Deep Learning Approach to Improve Skin Lesion Classification for Early Skin Cancer Detection," in *Proc. 2025 33rd Signal Process. Commun. Appl. Conf. (SIU)*, Istanbul, Turkey, 2025, pp. 1–4.
10. K. S. Sivaneshakumar, S. Vigneshwaran, K. G. Shanjay Kumar, N. Sathish Kumar, and A. Thejas Mecklinz, "Deep Learning-Based Classification and Detection of Skin Cancer," in *Proc. 2025 Int. Conf. Intell. Innov. Eng. Technol. (ICIET)*, Coimbatore, India, 2025, pp. 1–6.
11. N. Lakshmi Devi, V. V. R. Maheswara Rao, D. Vasanth Kumar, S. M. Padmaja, N. Silpa, and S. S. Reddy, "Deep Learning-Based Classification of Skin Lesions for Enhanced Dermatological Diagnosis," in *Proc. 2024 2nd Int. Conf. Netw., Multimed. Inf. Technol. (NMITCON)*, 2024, pp. 1–6.
12. T. Imran, A. S. Alghamdi, and M. S. Alkathairi, "Enhanced Skin Cancer Classification using Deep Learning and Nature-based Feature Optimization," *Eng., Technol. & Appl. Sci. Res.*, vol. 14, no. 1, pp. 12702–12710, Feb. 2024.
13. M. M. Yaqoob, S. Hamidullah, R. Manan, A. Khalid, U. F. Khattak, and M. A. Khan, "A Melanoma Skin Cancer Detection and Classification System Using Deep Transfer Learning," *Eng., Technol. & Appl. Sci. Res.*, vol. 16, no. 2, pp. 33570–33575, Jan. 2026.

-
14. M. Bekkaoui, H. Sebbagh, F. M. R. El Mansour, M. Merzoug, M. Hadjila, and M. M'hamedi, "Deep Learning-Based Melanoma Classification Using Hybrid DCNN-LSTM and DCNN-BILSTM Architectures," *Int. J. Comput. Exp. Sci. Eng. (IJCESEN)*, vol. 12, no. 2, pp. 474–485, May 2026.
 15. Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification With Deep Learning: A Systematic Review," *Front. Oncol.*, vol. 12, Art. no. 893972, Jul. 2022.
 16. P. Gaur and H. Sadawarti, "Deep Learning-Based Multi-Class Classification Of Skin Lesion Images," *Int. J. Drug Deliv. Technol. (IJDDT)*, vol. 16, no. 27s, pp. 601–617, Mar. 2026.
 17. M. A. O. Al-Yaqoubi, Y. N. Al Husaini, M. A. S. Al Husaini, M. H. Habaebi, W. AlKishri, and Y. A. K. Al Masqari, "Intelligent Framework for Skin Cancer Detection and Classification Using Deep Learning," in *Proc. 2026 IEEE 5th Int. Multidisciplinary Conf. Eng. Technol. (IMCET)*, Muscat, Oman, 2026, pp. 15–18.
 18. G. Dipesh, S. Eliyas, and S. Kumar, "Intelligent Skin Cancer Detection Using Hybrid Deep Learning and Machine Learning Model," in *Proc. 2026 3rd Int. Conf. Netw. Commun. (ICNWC)*, Padur, Chennai, India, 2026, pp. 1–7.
 19. N. Polu, S. A. Rafiya, G. S. Reddy, D. J. D. Srinivas, and K. S. B. Teja, "Skin Cancer Detection and Classification using Deep Learning," *J. Adv. Future Res. (JAAFR)*, vol. 4, no. 4, pp. 337–341, Apr. 2026.
 20. S. Khattar and T. Verma, "MobDensNet: Advanced Hybrid deep learning model for Multiclass Skin Cancer Classification," in *Proc. 2024 Int. Conf. Adv. Comput., Commun. Mater. (ICACCM)*, Gharuan, Mohali, India, 2024, pp. 1–5.
 21. N. Chorge and V. Parihar, "Skin Cancer Classification using Deep Learning and Explainable AI Techniques," in *Proc. 2025 IEEE Int. Conf.*, Pune, India, 2025, pp. 1–6.
 22. A. M. Sarhan, H. A. Ali, S. Yasser, M. Gobara, A. A. Kandil, G. Sherif, and E. Moustafa, "Achieving high-accuracy skin cancer classification with deep learning optimized by ant colony algorithm," *J. Electr. Syst. Inf. Technol.*, vol. 12, Art. no. 49, Jul. 2025.
 23. M. Watson *et al.*, "Multimodal models for skin cancer classification using clinical freetext and dermatoscopic images," *Commun. Med.*, vol. 6, Art. no. 277, Mar. 2026.
 24. K. R. Senthil Murugan, R. Deepalakshmi, A. S. V. Jayamaharaja, P. Balamurugan, and R. Vishal, "Skin Cancer Classification using Deep Learning Algorithms," in *Proc. 2023 Int. Conf. Data Sci., Agents, Artif. Intell. (ICDSAII)*, Madurai, India, 2023, pp. 1–5.
 25. R. N. Uma Mahesh, H. Vandana, G. H. Anvitha, V. G. Arpitha, and N. Khanum, "Skin Cancer Classification Using Deep Learning Techniques," in *Proc. 2025 Annu. Int. Conf. Data Sci., Mach. Learn. Blockchain Technol. (AICDMB)*, Mysuru, India, 2025, pp. 1–5.