Applied Data Science and Analysis
Vol.2026, **pp**. 111–126
DOI: https://doi.org/10.58496/ADSA/2026/007; ISSN: 3005-317X
https://mesopotamian.press/journals/index.php/ADSA

Research Article

DeepSkinNet: An EfficientNetV2B0-Based Framework for Binary Skin Lesion Classification

Zahraa Mohammed Hasan [1,*]

[1] *Directorate of Education in Najaf, Ministry of Education, Najaf, Iraq.*

ARTICLE INFO

Article History
Received 08 Jun 2026
Revised 20 Jul 2026
Accepted 09 Aug 2026
Published 14 Sep 2026

Keywords
Skin Lesion Classification,
Deep Learning,
EfficientNetV2B0,
Transfer Learning,
Medical Image Analysis,
DeepSkinNet,
HAM10000,
Test-Time Augmentation,
Validation-Optimized Decision Threshold.

ABSTRACT

The classification of skin lesions with deep learning is still difficult when there is a limited amount of labeled dermoscopic data. In this work, the authors propose an EfficientNetV2B0 framework for binary classification of actinic keratosis (AKIEC) and vascular lesions (VASC), which is referred to as DeepSkinNet. Images that were verified for this study were 270 images, which included 128 AKIEC images and 142 VASC images from the HAM10000 dataset. To avoid information leakage from multiple images in one dataset belonging to the same lesion, the datasets were split on a lesion-by-lesion basis using the HAM10000 lesion identifiers with the aim that no specific lesion would be present across training, validation, and test sets. The framework features two-stage selective fine-tuning, online data augmentation, class weighting, Test-Time Augmentation (TTA), and a decision threshold optimized for the validation set, as well as transfer learning with ImageNet. The optimized threshold (0.59) is solely based on the validation set and further applied to the independent test set. DeepSkinNet's performance on the lesion independent test set (44 images) was 93.18% accuracy, 92.86% balanced accuracy, 93.88% F1-score and 99.79% ROC-AUC. The findings in this study show that discrimination between AKIEC and VASC was promising under the protocol for the evaluation of leakage, and further validation on larger and external datasets is needed.

1. INTRODUCTION

Skin diseases are a major health problem all over the world, these lesions can be either benign or malignant. Early and correct diagnosis is crucial for the betterment of treatment and mortality. Dermoscopy is now a commonly used non-invasive imaging method for skin diagnosis and is highly dependent on the skill of dermatologists for interpretation. Therefore, automatic computer-aided diagnostic (CAD) systems have received more and more interest as a reliable tool to aid clinical decision making.The field of medical image analysis has witnessed remarkable achievements in the recent years due to the use of Artificial Intelligence (AI) and Deep Learning (DL) techniques, which allow automatic feature extraction directly from dermoscopic images. Classifying skin lesions using Convolutional Neural Networks (CNNs) has proven to be a challenging yet promising task with promising results, and the use of transfer learning has been widely adapted for the purpose of adapting a pre-trained model for a specific application, specifically skin lesion classification, while with smaller training set sizes, learning cost and convergence speed are improved [2], [3]. In recent deep learning architectures, EfficientNetV2 is a suitable backbone, using the EfficientNet family, and is training-aware in terms of its architectural design [4]. In this work, the pretrained backbone for the proposed classification framework was chosen as EfficientNetV2B0. But the classification of the skin lesions is still a challenging task, because of the limited number of labeled data, class imbalance, visual similarity between categories of lesions and variations in dermoscopic images. Further, when several images of the same lesion are split between the training and test sets, evaluation methods using image-level random splitting might risk information leakage. Based on this, this study aims to present a new framework, named DeepSkinNet, which is based on the efficient network EfficientNetV2B0 for a binary classification problem for Actinic Keratosis (AKIEC) and Vascular Lesions (VASC). The proposed pipeline combines transfer learning, selective fine-tuning, online data augmentation, class weighting, Test-Time Augmentation (TTA), and a validation set optimized decision threshold. A verified binary subset of the HAM10000 dataset was used, and lesion-level partitioning was followed to avoid having images of the same lesion being distributed across the training/validation/test subsets. The accuracy, balanced

*Corresponding author. Email: zmhmdhsn067@gmail.com

accuracy, F1 and ROC-AUC of DeepSkinNet on the independent test set were 93.18%, 92.86%, 93.88% and 99.79% respectively under this leakage controlled evaluation protocol.

The main contributions of this work are summarized as follows:

1. Design of DeepSkinNet: EfficientNetV2B0 based framework for binary classification of Actinic Keratosis (AKIEC), Vascular Lesion (VASC).
2. Implementation of a lesion-level leakage controlled partitioning methodology with HAM10000 lesion identifiers to avoid leakage of images of the same lesion across the three datasets (training, validation, test).
3. Incorporation of transfer learning/ fine-tuning/ online data augmentation/ class weighting into the training pipeline.
4. Use of Test-Time Augmentation (TTA) [12] and a validation-optimum decision threshold determined solely from the validation set, but applied in an independent test set.

Evaluation with complementary metrics such as the accuracy, precision, recall, specificity, F1-score, balanced accuracy and ROC-AUC.

2. RELATED WORK

2.1 Deep Learning for Skin Lesion Classification

The effectiveness of acquiring discriminative features directly from dermoscopic images from the images, without handcrafted feature extraction has made deep learning the preferred method for automated classification of skin lesions. Convolutional Neural Networks (CNNs) have been shown to outperform other methods in categorizing visually similar categories of skin lesions, providing a much needed advance in computer-aided diagnosis.Transfer learning has also been applied to pretrained models like ResNet[6], DenseNet[7], MobileNetV2 [8], EfficientNet, and Vision Transformers (ViTs) [11]to further boost classification performance by increasing feature generalization and decreasing the reliance on large annotated medical datasets. These approaches have become the standard measures for publicly available datasets such as HAM10000 [1] and ISIC.However, these advances have not overcome other problems, including the small number of images in the dataset, class imbalance, lesion similarity, and imaging variability, which make classification unreliable. As a result, recent studies have been dedicated to developing more powerful and efficient deep learning frameworks for enhancing diagnosis accuracy and practical clinical applications.

2.2 Transfer Learning in Skin Lesion Analysis

Transfer learning is now one of the most powerful techniques for medical image classification, for instance when big annotated datasets are unavailable. Transfer learning is the process of taking advantage of already learned representations from a large scale image dataset like ImageNet [5] and fine-tuning them to a medical imaging task. This approach will not only save a lot of compute resources, but it will also be faster to train and converge the model.Transfer learning is a technique that has been extensively used in the field of skin lesion classification due to its size limitation and class imbalance of dermoscopic datasets. Researchers have shown that using pretrained feature extractors allows them to achieve significant gains in classification accuracy without overfitting. The fine-tuning of selected layers of pre-trained networks has also increased the amount of feature representation by enabling the model to learn task-specific features of skin lesions while retaining general visual knowledge gained during pretraining.Many CNN architectures have been used with transfer learning with success: ResNet, DenseNet, Inception, MobileNet and EfficientNet. From comparative studies, it is observed that none of the architectures is always superior to others for different datasets; it is very much dependent on the characteristic of the datasets, the preprocessing technique, the optimization technique, and the training method. In view of this, the research has turned to the combination of transfer learning with other complementary approaches like sophisticated data augmentation, adaptive optimization, ensemble learning and uncertainty-aware inference to enhance the diagnostic accuracy.Many transfer learning-based methods continue to use traditional prediction pipelines with a one-pass inference and a fixed decision threshold. These strategies might compromise the robustness of classification, especially for challenging dermoscopic images. Therefore, the use of adaptive prediction mechanisms along with transfer learning has emerged as a new trend to enhance the reliability of such automated skin lesion classification systems.

2.3 EfficientNet-Based Approaches

To achieve efficient scaling of network depth, width and input resolution, EfficientNet proposed a compound scaling method [3]. EfficientNetV2 is the next improvement of this family, which is trained-aware in its architecture design and progressive learning [4]. Skin lesion classification has been explored using efficientNet based architectures in transfer learning. In the current study, the proposed DeepSkinNet framework is based on the EfficientNetV2B0 pretrained backbones. In this study an independent computational-efficiency benchmark to alternative architectures was not performed. To implement Test-Time Augmentation (TTA), predictions were made using random transformations of each input image, and then the results were averaged. Furthermore, the decision threshold was tuned solely on the validation data, and then set in a fixed manner for the test evaluation. The proposed DeepSkinNet framework adopted transfer learning, selective fine-tuning, data augmentation and class-weight balancing.

3. MATERIALS AND METHODS

The methodology used for the proposed DeepSkinNet framework for binary skin lesion classification is described in this section. It adopts EfficientNetV2B0 architecture, transfer learning, selective fine-tuning, data augmentation, class weighting, Test-Time Augmentation (TTA) and decision threshold selection by validation. The overall workflow begins with the acquisition of dermoscopic images from the HAM10000 dataset [1] followed by the verification of the metadata, preprocessing of the images and partitioning of the images at the level of the lesions. The online data augmentation technique is applied to the training subset during training, and a pretrained EfficientNetV2B0 model is used as the backbone feature extractor with selective fine-tuning. In inference, both TTA and a threshold value based solely on the validation set are applied. Lastly, the proposed framework is tested with Accuracy, Precision, Recall, F1-score, Specificity, Balanced Accuracy and ROC-AUC.

3.1 Overall Workflow

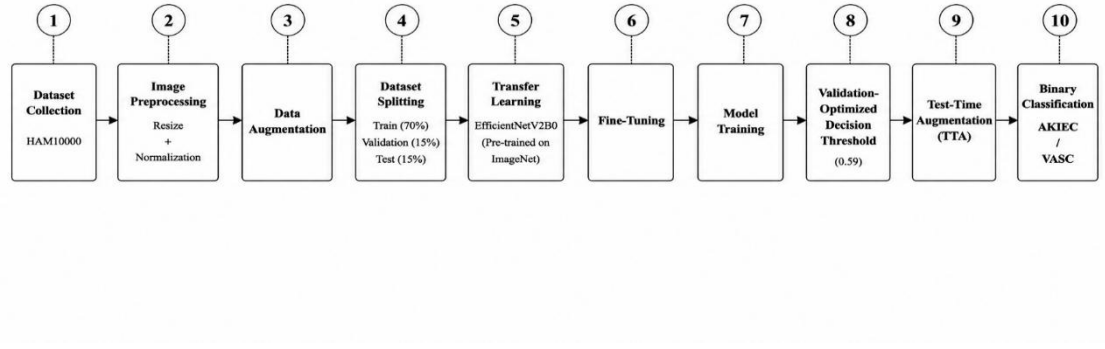

Fig. 1. The overall flow of the proposed DeepSkinNet system

The initial step is to use the image set HAM10000 [1] followed by metadata verification, image preprocessing, and partitioning of lesions into training, validation, and testing sets. Online data augmentation only happens to the training set. Finally, transfer learning is implemented with EfficientNetV2B0 as the backbone network and then selective fine-tuning is performed for such a binary classification task as skin lesion classification. Inference is performed using Test-Time Augmentation (TTA) [12] and the final AKIEC–VASC classification is based on the decision threshold chosen exclusively on the basis of the validation set, which is then fixed for test evaluation.

3.2 Dataset

In this study, experiments were carried out with a binary subset of the publicly available HAM10000 [1] dataset, which is made up of dermoscopic images of 10015 images belonging to seven diagnostic categories. The present study dealt with two classes: Actinic Keratosis (AKIEC) and Vascular Lesion (VASC), which was a binary classification problem For

HAM10000, the two classes chosen to define a focused binary proof-of-concept classification task enabled a focused evaluation of the proposed pipeline, and a leakage-controlled lesion-level evaluation protocol, rather than trying to perform a complete multiclass diagnosis.A total of 272 candidate images were initially collected (130 AKIEC and 142 VASC). The images were checked against the official images metadata HAM10000 by their image identifiers. From the 272 candidate images, 270 images were found and retained to have a valid corresponding lesion identifier (lesion_id), and included in leakage-controlled analysis (128 images were identified as AKIEC and 142 were identified as VASC). Due to the fact of the two images being unable to be found in HAM10000 there was no matching with the metadata, therefore the two AKIEC images were not included in the revised experiment.An image-level random split might also lead to images from the same lesion being in different subsets, introducing information leakage, as HAM10000 contains multiple dermoscopic images of the same lesion. Parting of the datasets was achieved at the level of the lesion_id with a fixed random seed of 42 and stratification per lesion class. About 70% of the data was set aside for training and the remaining data split into validation and testing subsets. A total of 187, 39, and 44 images were used for training, validation, and testing, representing 129, 28, and 28 unique lesions. The lesion identifiers were checked to be mutually exclusive between the 3 subsets to ensure there were no overlapping lesions within each of the 3 subsets.

TABLE I. DISTRIBUTION OF THE LEAKAGE-CONTROLLED LESION-LEVEL DATASET ACROSS TRAINING, VALIDATION, AND TESTING SUBSETS.

Dataset Subset	AKIEC	VASC	Total
Training	90	97	**187**
Validation	17	22	**39**
Testing	21	23	**44**
Total	**128**	**142**	**270**

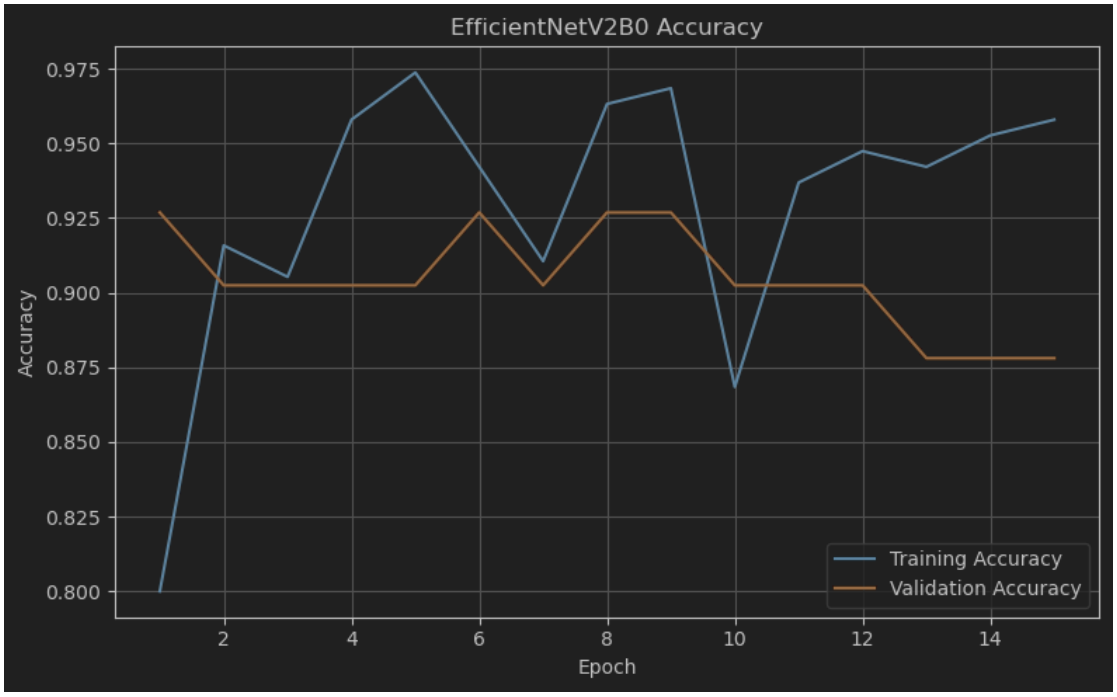

Fig. 2. shows the training and validation accuracy curves across the training epochs. These curves are presented to illustrate the optimization behavior of the model during training and validation.

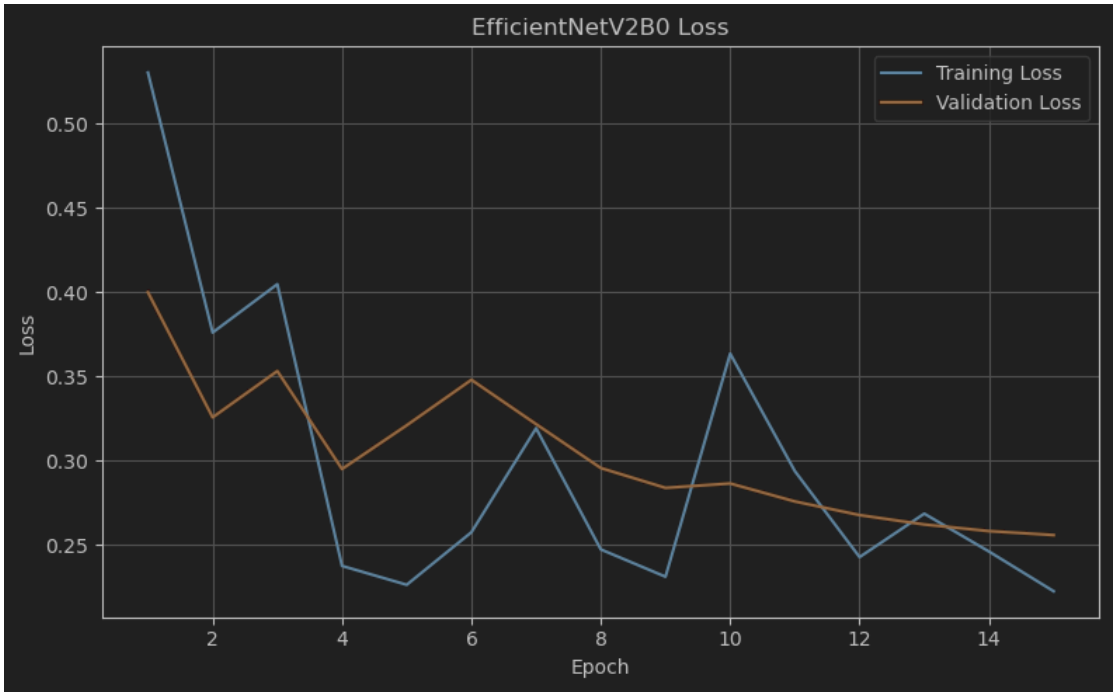

Fig. 3. shows the training and validation loss curves across the training epochs. These curves are presented to illustrate the loss behavior during model optimization.

3.3 Image Preprocessing

To train the models, all dermoscopic images were preprocessed using the same pipeline. The images used from the subset of the HAM10000 dataset [1] were resized to 224 × 224 pixels without losing the RGB color format to comply with the input requirements of EfficientNetV2B0. The input preprocessing applied to pixel values was the same throughout the processing that was used by the EfficientNetV2B0 pipeline. The lesion categories were converted into binary labels for training and testing of the model. To ensure that the images of the same lesion were not split between the training, validation and testing subsets, the data set was partitioned at the lesion level as described in Section 3.2. Data augmentation was only performed during training, as outlined in Section 3.4. This uniform pre-processing approacensured uniformity in subsequent transfer learning and model optimization.

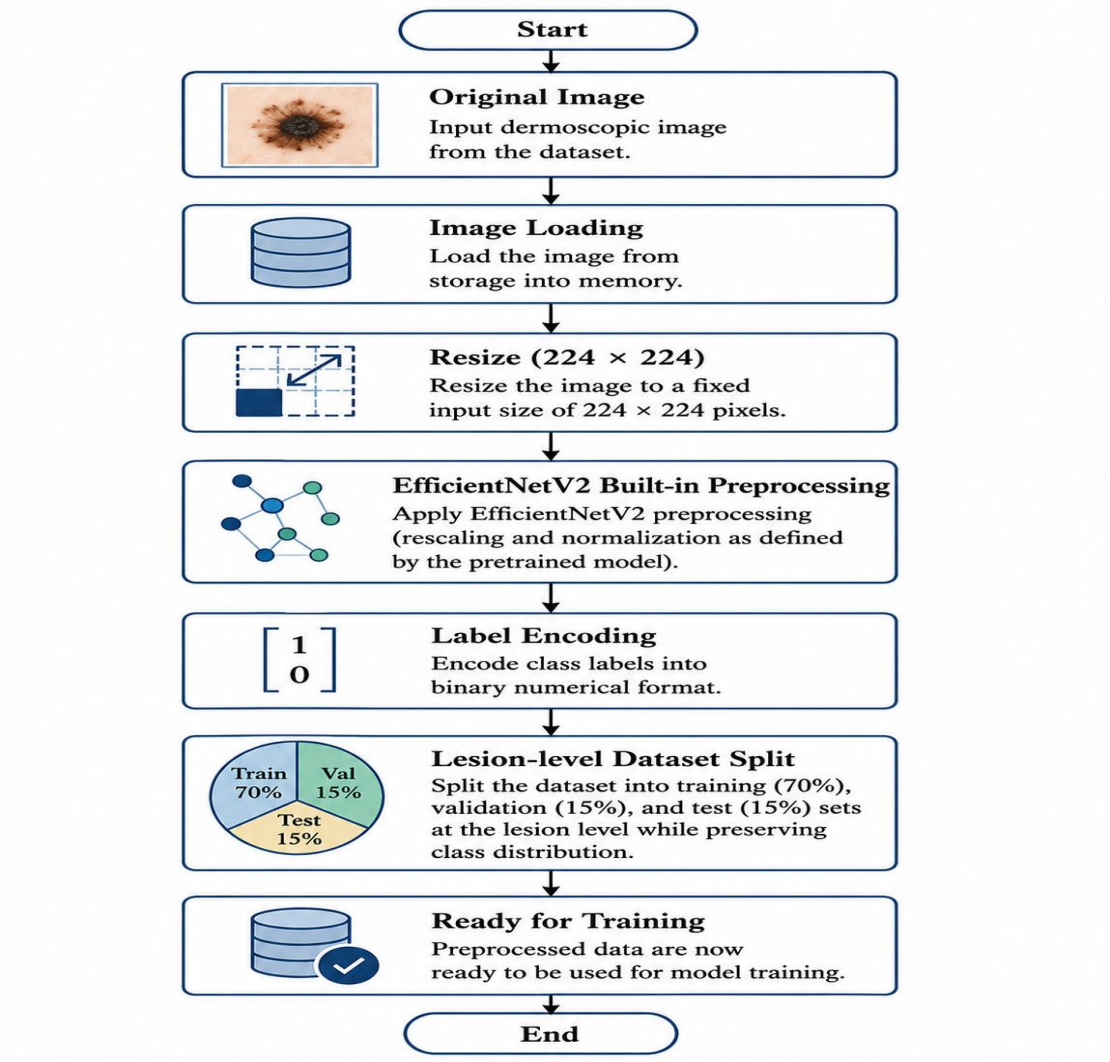

Fig. 4. Data Preprocessing Workflow of the Proposed DeepSkinNet Framework

3.4 Data Augmentation

Online data augmentation was used on the dermoscopic data during the training of the network.This data augmentation technique was applied in the online training of the network, to avoid overfitting due to the limited size of the dermoscopic dataset. The augmentation pipeline featured random horizontal and vertical flips, random rotations (by 0.15), random zooming (by 0.15), random contrast adjustment (by 0.15), and random translation (height: 0.08, width: 0.08). These mutations were not created as extra pictures but as a dynamic result of the process of training. Thus, the size of the training set was not changed and the model was presented with different representations of the training samples. The data augmentation was limited to during training and not used to artificially increase the size of the validation and testing subsets. This augmentation strategy was used to increase the diversity of image transformations presented to the model during training.

3.5 Proposed DeepSkinNet Architecture

The proposed DeepSkinNet is designed with the EfficientNetV2B0 model for binary skin lesion classification framework. The backbone feature extractor used was an EfficientNetV2B0 model which was pre-trained on ImageNet [5] and the top classification layer was removed. The built-in efficientNetV2 preprocessing in the backbone input pipeline was used in the framework.After feature extraction, the backbone output was passed through a Global Average Pooling layer and then a Batch Normalization and a Dropout layer [10] (rate=0.40). This was followed by a fully connected Dense layer with 256 units and Swish activation, and L2 kernel regularization (1×10^{-4}). To further reduce over-fitting, a second Dropout layer

with a rate of 0.30 was added. Finally, a single sigmoid output neuron output probability for binary classification.During inference, a trained model, Test-Time Augmentation (TTA) [١2] and a decision threshold optimized on the validation set were used. The details of the training and fine-tuning strategy are given in Section 3.6 and the inference procedure is given in Section 3.7.

TABLE II. ARCHITECTURE OF THE PROPOSED DEEPSKINNET FRAMEWORK

Stage	Layer / Operation	Output / Setting	Purpose
Input	RGB Dermoscopic Image	224 × 224 × 3	Model input
Augmentation	Online augmentation pipeline	Training only	Improve training diversity
Backbone	EfficientNetV2B0 pretrained on ImageNet	Feature maps	Deep feature extraction
Pooling	Global Average Pooling	Feature vector	Reduce spatial dimensionality
Normalization	Batch Normalization	—	Stabilize feature representation
Regularization	Dropout	0.40	Reduce overfitting
Classifier	Dense + Swish	256 units	High-level feature learning
Regularization	L2 kernel regularization	1×10^{-4}	Control model complexity
Regularization	Dropout	0.30	Reduce overfitting
Output	Dense + Sigmoid	1 unit	Binary probability prediction
Decision	Validation-optimized threshold	0.59	Convert probability to class label
Inference	Test-Time Augmentation (TTA)	Final probability	Aggregate predictions across predefined transformations

3.6 Training Strategy

A two-stage transfer learning approach was used for training the proposed DeepSkinNet framework. The EfficientNetV2B0 backbone was frozen in the first stage, and only the new classification head was trained. The Adam optimizer [9] and binary cross-entropy loss function were used in this stage with a learning rate of 1×10^{-3} and 15 epochs max. The unfrozen and

fine-tuned upper layers of the EfficientNetV2B0 backbone were trained in the second stage with a smaller learning rate of 1×10^{-5} for up to 15 more epochs. For consistency, training used a fixed random seed of 42, and a batch size of 8.To mitigate the impact of the moderate imbalance in classes, class weighting was used in training. ModelCheckpoint was set up to track the AUC of the validation set and to keep the checkpoint with the best performance. Early stopping was also used to stop the model when validation AUC was not improving after 5 epochs and to retrieve the best model weights. Moreover, ReduceLROnPlateau tracked the validation loss with a reduction factor of 0.25, a patience of 2 epochs and a minimum learning rate of 1×10^{-7}. These procedures were used to manage the optimization process and to determine when training should be stopped based on validation performance. To account for the remaining class imbalance in the training subset, class weights were calculated using the balanced weighting formula wc=N/(K×nc)w_c = N/(K \times n_c), where NN is the total number of training images, KK is the number of classes, and ncn_c is the number of training images belonging to class cc. With N=187N = 187, K=2K = 2, nAKIEC=90n_{AKIEC} = 90, and nVASC=97n_{VASC} = 97, the resulting class weights were 1.0389 for AKIEC and 0.9639 for VASC, respectively. These weights were applied only during model training.

3.7 Prediction Strategy

The proposed DeepSkinNet framework introduces a combination of Test-Time Augmentation (TTA) [12] and a validation-optimized decision threshold with the validation-guided prediction strategy. The prediction probabilities are aggregated with TTA instead of using a single forward pass with the standard threshold of 0.5, the final threshold is determined only with the help of the validation set. The threshold is then set and used for the independent test set. This is a procedure that does not utilize test-set information during the selection of thresholds, and gives a reproducible inference procedure.

3.7.1 Test-Time Augmentation

During validation and testing, Test-Time Augmentation (TTA) [12] was used to generate multiple transformed versions of images to get prediction probabilities for each image. The six versions tested were: Original image, horizontal flip, vertical flip, 90°, 180°, and 270° rotations. The probability was produced by the trained model for each transformed image and the mean of the six predictions was calculated to get the final probability. This mean-aggregation procedure was used consistently for both the validation and test sets. TTA was used to decrease the sensitivity towards image orientation and geometric variation without altering the trained model parameters.

3.7.2Validation-Optimized Decision Threshold

Obtained the TTA-averaged prediction probabilities and optimized the decision threshold only on validation set. Different thresholds (between 0.20 and 0.80) were tested at intervals of 0.01. Balanced accuracy, Accuracy and F1 score were computed for each candidate threshold. The thresholds for candidates were prioritized mainly by Balanced Accuracy (and in the event of ties, by Accuracy and F1-score). The validation optimized threshold was 0.59 based on this procedure. This value was subsequently set and then applied to the independent test set without changes. Thus, the threshold is not adaptive, but rather validation-optimized, meaning that it does not change for individual test samples.

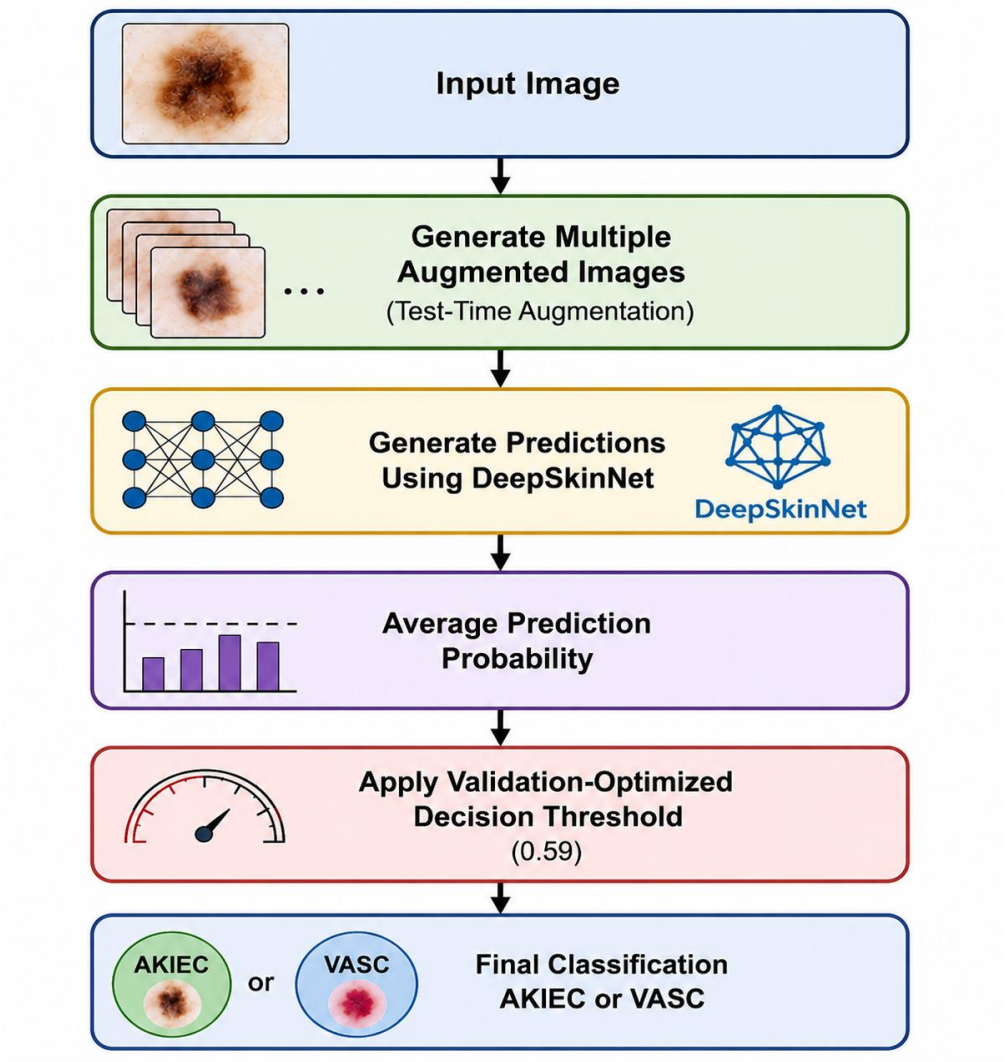

Fig. 5. Validation-Guided Prediction Pipeline of DeepSkinNet

3.8 Evaluation Metrics

The performance of the proposed DeepSkinNet framework has been quantitatively evaluated by employing various metrics. These included Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, Balanced Accuracy and ROC-AUC. Complementary metrics were also reported as a way to capture class-specific performance, given that overall accuracy does not necessarily describe performance on each class, especially when there is class imbalance.In the binary evaluation, Vascular Lesion (VASC) was coded as label 1 and therefore considered positive class, while Actinic Keratosis (AKIEC) was coded as label 0 and considered as the negative class. The binary Precision, Recall (Sensitivity) and Specificity were therefore reported based on VASC as the positive class. To prevent the model performance from being misinterpreted, an additional analysis by class of the precision, recall and F1-score was conducted for both categories of lesions.Accuracy – percentage of correctly classified test images. Sensitivity refers to the percentage of VASC images that are correctly classified, while Specificity refers to the percentage of AKIEC images that are correctly classified, according to this label convention. Balanced Accuracy is the average recall over the two classes. The model performance in terms of its discriminative power was measured using the ROC-AUC, which is not dependent on the final selected classification threshold. All final performance metrics were calculated on the lesion-independent test set after the decision threshold was chosen only based on the performance on the validation set.

4.1 EXPERIMENTAL RESULTS

The experimental analysis of DeepSkinNet is presented in this section. The aim of the experiments was to evaluate the performance of the proposed DeepSkinNet classification framework on the binary version of the HAM10000 [1] database. To ensure an extensive assessment of the capability of classification, the model was evaluated on an independent testing set, and various quantitative evaluation metrics were used.

4.1 Experimental Environment

The experiments were performed on a 64-bit Windows 10 platform with Python 3.9.13 and the Keras API and TensorFlow 2.20.0. Experiments were run without GPUs, with an Intel64 Family 6 Model 126 CPU.A pretrained backbone network, efficientnetv2b0 trained on ImageNet [5] was chosen. After metadata verification, 270 dermoscopic images were used in the updated experiments, 128 of which were Actinic Keratosis (AKIEC) and 142 of which were Vascular Lesion (VASC). To avoid lesion-level information leakage, the data was divided at the lesion level into training, validation, and independent testing sets as described in Section 3.2. The sizes of the resulting subsets were 187, 39 and 44 images, respectively .This two-stage transfer learning approach (see Section 3.6) was combined with online data augmentation, class weighting, early stopping, learning-rate scheduling, model checkpointing, Test-Time Augmentation (TTA) [12] and validation-optimized decision threshold (see Section 3.7) in the experimental pipeline.

4.2 Training Performance

The training process showed for stable convergence during optimization process. The classification head was easily trained to obtain discriminative representations while using the pre-trained feature extractor EfficientNetV2B0 during the first transfer learning phase. Both the training and the validation performances were found to be improved after selective fine-tuning while there was not remarkable overfitting.The best model was automatically saved during the optimization process with the help of the ModelCheckpoint mechanism, and this was monitored by the validation dataset. To overcome convergence instability from the early stopping, adaptive learning-rate scheduling was additionally adopted. The decision threshold was optimized exclusively on the validation set using the procedure described in Section 3.7.2. The resulting validation-optimized decision threshold was 0.59 and was subsequently fixed and applied to the independent testing set without further adjustment.. During the two training periods, the training history was monitored for model optimization. The results of the individual contributions of transfer learning, selective fine tuning, learning-rate scheduling, and other pipeline components to the final performance cannot be separated from the present experiment, however, as a controlled ablation study was not performed.

TABLE III. FINAL CLASSIFICATION PERFORMANCE

Metric	Value
Accuracy	**93.18%**
Balanced Accuracy	**92.86%**
Precision	**88.46%**
Recall (Sensitivity)	**100.00%**
Specificity	**85.71%**
F1-score	**93.88%**
ROC-AUC	**99.79 %**
Validation-Optimized Threshold	**0.59**

Table 3 shows the final performance of the proposed DeepSkinNet framework on the lesion independent test set. The overall accuracy of the model was 93.18% and the balanced accuracy was 92.86%. The model had a precision of 88.46%, sensitivity (recall) of 100.00%, specificity of 85.71%, and an F1-score of 93.88% with VASC as a positive class. The ROC-AUC is 99.79%, with a good discriminative ability over the different classification thresholds. The final classification results were derived with a decision threshold value of 0.59, as this was chosen only on the validation set and then applied to the independent test set.

4.3 Classification Performance

In addition to overall accuracy, several complementary performance metrics were used to comprehensively evaluate the binary classification performance of the proposed DeepSkinNet framework, such as Precision, Recall (Sensitivity), Specificity, F1-score, Balanced Accuracy, and ROC-AUC. VASC was treated as positive class (label = 1) and the class AKIEC was treated as the negative class (label = 0).The balanced accuracy is 92.86%, while the F1-score for the positive class VASC is 93.88%, reflecting the balance between precision and recall for the VASC class. Furthermore, the ROC-AUC of 99.79% suggests a good discrimination between the two lesion categories for all the decision thresholds. The confusion matrix is further discussed in Section 4.4 and the performance within each class is discussed in Section 4.5.

4.4 Confusion Matrix Analysis

The confusion matrix is used to get a detailed view of the classification result by showing how many samples were correctly and incorrectly classified in each of the categories of lesions. Overall accuracy is useful for summarizing model performance but the analysis here provides more insight, by providing the distribution of true positive, true negative, false positive, and false negative predictions.The experimental results presented here confirm that the proposed DeepSkinNet framework has learned discriminative visual features of both Actinic Keratosis and Vascular Lesions and most of the dermoscopic images were classified correctly. The misclassification rate was low with only a few samples misclassified, as this could happen because some dermoscopic patterns may look alike.Minimising the number of false-negative predictions is especially significant from a clinical point of view, since the absence of suspicious lesions may result in the delay of diagnosis and treatmentThe confusion matrix gives a class-wise summary of the classification results obtained for the lesion-independent test set, and is to be used in conjunction with the quantitative performance parameters discussed in Table 3.

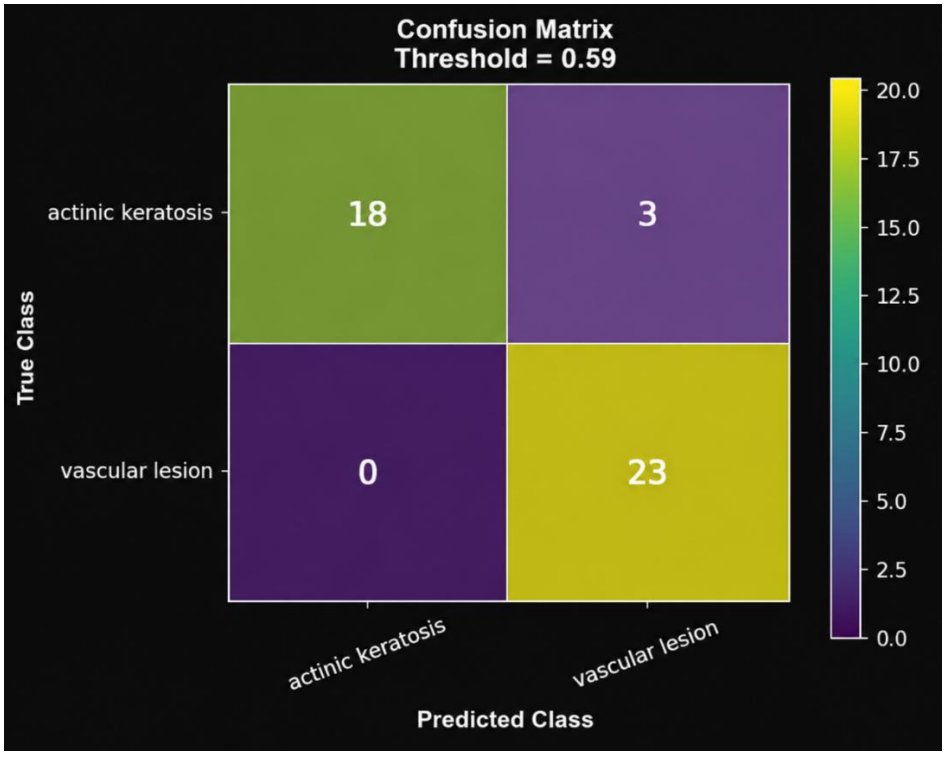

Fig. 6. Confusion matrix of the proposed DeepSkinNet framework on the lesion-independent test set.

The confusion matrix from the lesion independent test set at the validation optimised decision threshold of 0.59 is shown in Figure 6. There were 21 AKIEC images among which 18 were correctly classified as AKIEC and 3 as VASC. 23 of the images were correctly identified as VASC while no VASC image was misclassified as AKIEC. As such, the model was able to obtain a class-specific recall rate of 85.71% for AKIEC and 100.00% for VASC. The results are in line with the convention of the positive class as outlined in Section 3.8, in which the positive class is defined as the VASC.

4.5 ROC Curve Analysis

The effectiveness of the proposed DeepSkinNet framework for discrimination ability under various classification thresholds was evaluated using Receiver Operating Characteristic (ROC) analysis. The ROC curve shows the probability of a true positive (TP) across a range of decision thresholds and the probability of a false positive (FP). On the lesion-independent test set, the proposed framework obtained the ROC-AUC of 0.9979 (99.79%) following Test-Time Augmentation (TTA), as demonstrated in Figure 7. The high ROC-AUC suggests good discrimination ability between Actinic Keratosis (AKIEC) and Vascular Lesion (VASC) in the test set assessed. Importantly, ROC-AUC is threshold independent, meaning that the decision threshold that was optimized for validation was not used for calculating ROC-AUC, but only to get final binary class predictions.

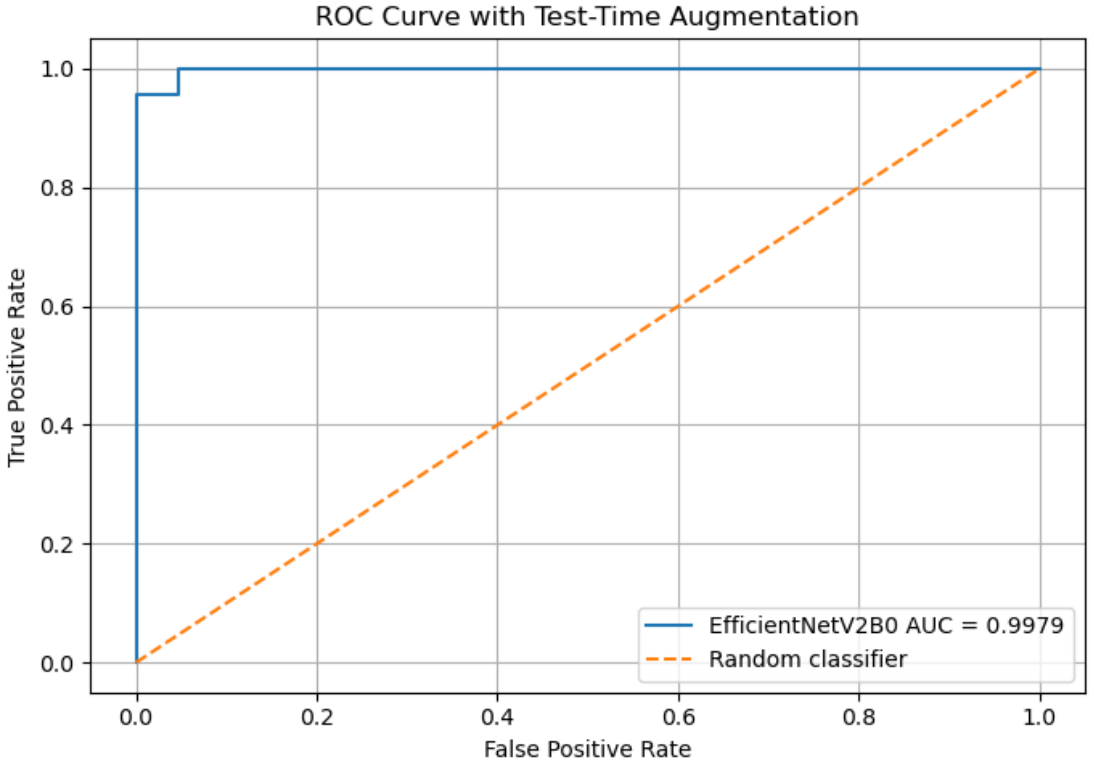

Fig. 7. Receiver Operating Characteristic (ROC) curve of the proposed DeepSkinNet framework on the lesion-independent test set.

The TTA-averaged prediction probabilities were used to obtain the Receiver Operating Characteristic (ROC) curve on the test set comprised of lesions that were not used to train the TTA. There was good discrimination between AKIEC and VASC with a ROC-AUC of 0.9979 (99.79%) for the model for all decision thresholds. The value of ROC-AUC is threshold independent, and represents the rankability of the model rather than the threshold used in the validation process (0.59). Since a controlled ablation analysis was not performed, the contributions of TTA to the performance could not be separated out from the current experiment, although it was included in the inference pipeline.

4.6 Qualitative Analysis and Comparison with Recent Studies

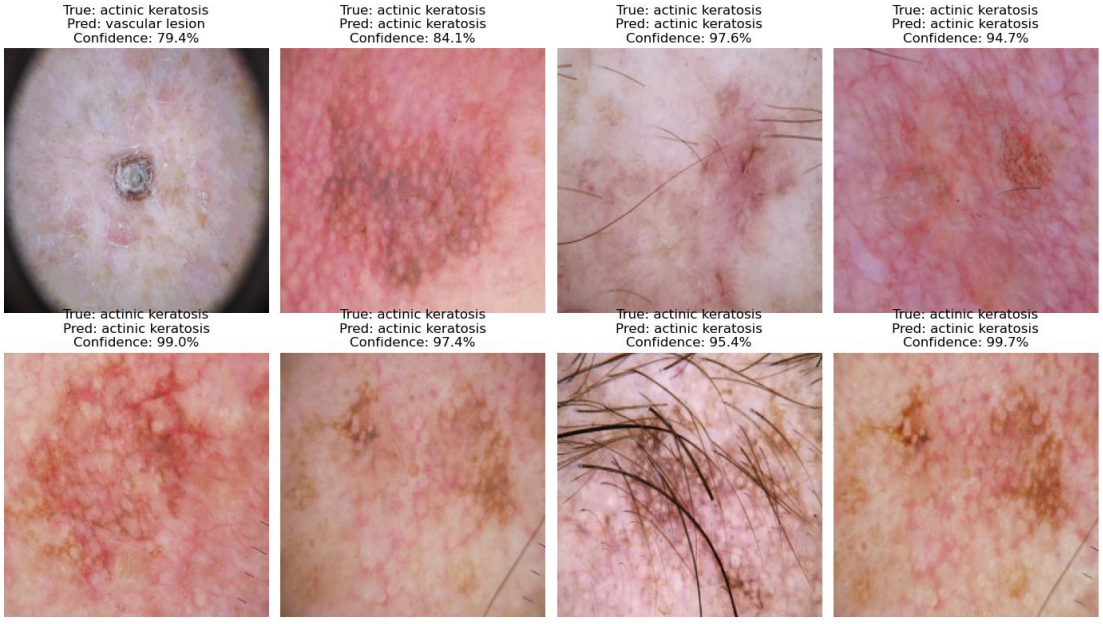

Fig. 8. Representative prediction examples generated by the proposed DeepSkinNet framework on the testing dataset.

In Figure 8, some representative prediction examples for the lesion-independent test set are displayed. For each example, the ground truth class, the predicted class, and the confidence score are shown. The overall performance is satisfactory, with most of the samples correctly classified, and a few samples that are misclassified representing the remaining challenges in distinguishing visually similar lesion patterns. The qualitative examples are intended to be used as a supplement to the quantitative assessment and are not intended to be considered as evidence of clinical validation.

TABLE IV. COMPARISON BETWEEN THE PROPOSED DEEPSKINNET FRAMEWORK AND RECENT STUDIES

Study	Year	Dataset	Classification Task	Model / Method	Accuracy (%)	ROC-AUC (%)
Ali et al. [16]	2022	HAM10000	Seven-class	EfficientNet-B4 with transfer learning and fine-tuning	87.91	NR
Qian et al. [17]	2022	HAM10000	Multiclass	CNN with grouped multi-scale attention and class-specific loss weighting	91.60	97.10
Choi et al. [18]	2024	HAM10000	Multiclass	ABC-based preprocessing	93.26	NR
Hu et al. [14]	2024	HAM10000	Multiclass	EfficientNetV2-based multi-scale feature fusion and class-weight loss	94.00	99.30

Study	Year	Dataset	Classification Task	Model / Method	Accuracy (%)	ROC-AUC (%)
Behara et al. [15]	2024	HAM10000 and ISIC 2020	Skin lesion classification	Active contour segmentation, ResNet50, and attention-guided Capsule Network	98.00	97.30
Das and Addya [19]	2025	HAM10000	Seven-class	EfficientNet-B0 with fine-tuning, TTA, and Monte Carlo Dropout	97.15	NR
Proposed DeepSkinNet	**2026**	HAM10000 binary subset: AKIEC and VASC	Binary	EfficientNetV2B0, transfer learning, selective fine-tuning, class weighting, TTA, andvalidation-optimized decision threshold	93.18	99.79

Table 4 provides a summary of some recent representative studies which have tested deep learning methods for skin lesion diagnosis based on HAM10000 or similar dermoscopic datasets. It is not possible to directly compare reported results as the studies vary in classification tasks, number of classes, composition of datasets, preprocessing, data-partitioning protocols, and evaluation settings. In particular, the proposed DeepSkinNet framework is based on a binary AKIEC–VASC classification task with a split performed at the level of the lesions, while the majority of the listed works are based on multiclass classification tasks. The table is therefore presented not to assert that the methodology used is superior to that of previous works, but rather to put the methodological characteristics of the proposed framework in perspective.

4.7 Limitations of the Proposed Framework

However, the proposed DeepSkinNet framework has the following caveats that should be taken into account when interpreting the results: The experimental assessment was done on a small subset of the HAM10000 dataset [1] consisting of 270 images of the two classes (Actinic Keratosis (AKIEC) and Vascular Lesion (VASC)) identified from 185 different lesions, metadata verified. The images from the same lesion were not appearing in the training, validation and test sets, but there were only 44 images from 28 unique lesions in the final test set. So that, it should be used with caution and care in reporting the performance. Second, the evaluation used only one data split for each lesion, and there was no external independent data set used for evaluation. Third, a controlled ablation study was not conducted, so that the isolation of the influence of each of the above factors is not feasible in the present experiment. Furthermore, the model calibration and explainability techniques, such as Grad-CAM [١٣] or SHAPwere not tested. However, future studies should focus on increasing the number and diversity of the datasets, external validation, multiclass classification, controlled ablation experiments, calibration evaluation and explainability analysis before the broader clinical applicability.

5. DISCUSSION

The experimental results show the promising results of the proposed DeepSkinNet framework for binary classification of Actinic Keratosis (AKIEC) and Vascular Lesion (VASC) on the test set without any lesions. The framework performed with an accuracy of 93.18%, balanced accuracy of 92.86%, F1-score of 93.88% and ROC-AUC of 99.79% using the validation optimized decision threshold of 0.59. The confusion matrix demonstrated that all 23 VASC images were correctly classified, and 18 of 21 AKIEC images were correctly classified with 3 being misclassified as VASC. These results suggest that there was a relatively strong discrimination between the two classes probed in the present experimental context.The pipeline proposed in this work involves transfer learning, class-weight balancing, Test-Time Augmentation (TTA) [١2] and threshold selection via validation. The contribution of each component to the overall performance cannot be measured, however, since a controlled ablation analysis was not performed. Furthermore, the relatively small test set of lesions with lesion-independence limits generalizability. The framework needs to be further evaluated on larger datasets, external datasets, multi-class settings, and ablation studies with controlled settings as well as methods that explain the model like Grad-CAM or SHAP to determine the robustness and wider applicability of the framework.

6. CONCLUSION

This study involved an evaluation of the DeepSkinNet framework, which is based on the EfficientNetV2B0 network architecture, in the task of binary skin lesion classification using a subset of images from HAM10000 [1] database that are metadata-verified as Actinic Keratosis (AKIEC) and Vascular Lesion (VASC). The proposed framework involves the use of transfer learning, along with selective fine-tuning, class-weight balancing, Test-Time Augmentation (TTA) [12] and validation-based decision threshold selection. To prevent data leakage, the data was divided on the lesion level, with images from the same lesion not appearing in all three sets (training, validation, and testing).Given the validation optimized decision threshold of 0.59, DeepSkinNet obtained an accuracy of 93.18%, balanced accuracy of 92.86%, F1-score of 93.88% and ROC-AUC of 99.79% on the lesion-independent test set. These findings are an encouraging discrimination between AKIEC and VASC in the present experimental conditions. The relatively small dataset and test set, however, and the lack of external validation and ablation analysis limit the generalizability of the findings. The framework needs to be assessed on larger, independent data sets and needs to be expanded to multiclass skin lesion classification, and methods for calibration and explainability should be explored in future work to determine its potential for broader clinical use.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

None.

References

[1] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” Scientific Data, vol. 5, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[3] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.

[4] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 10096–10106.

[5] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.

[11] K. Nakai, Y.-W. Chen, and X.-H. Han, "Enhanced deep bottleneck transformer model for skin lesion classification," *Biomedical Signal Processing and Control*, vol. 78, Art. no. 103997, 2022, doi: 10.1016/j.bspc.2022.103997.

[12] A. T. Ibrahim, M. Abdullahi, A. F. D. Kana, M. T. Mohammed, and I. H. Hassan, "Categorical classification of skin cancer using a weighted ensemble of transfer learning with test-time augmentation," *Data Science and Management*, vol. 8, no. 2, pp. 174–184, 2025, doi: 10.1016/j.dsm.2024.10.002.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[14] Z. Hu, W. Mei, H. Chen, and W. Hou, "Multi-scale feature fusion and class weight loss for skin lesion classification," Computers in Biology and Medicine, vol. 176, Art. no. 108594, 2024, doi: 10.1016/j.compbiomed.2024.108594.

[15] K. Behara, E. Bhero, and J. T. Agee, "An Improved Skin Lesion Classification Using a Hybrid Approach with Active Contour Snake Model and Lightweight Attention-Guided Capsule Networks," Diagnostics, vol. 14, no. 6, Art. no. 636, 2024, doi: 10.3390/diagnostics14060636.

[16] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer," Neuroscience Informatics, vol. 2, no. 4, p. 100034, 2022, doi: 10.1016/j.neuri.2021.100034.

[17] S. Qian, K. Ren, W. Zhang, and H. Ning, "Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting," Computer Methods and Programs in Biomedicine, vol. 226, p. 107166, 2022, doi: 10.1016/j.cmpb.2022.107166.

[18] J.-Y. Choi, M.-J. Song, and Y.-J. Shin, "Enhancing Skin Lesion Classification Performance with the ABC Ensemble Model," Applied Sciences, vol. 14, no. 22, Art. no. 10294, 2024, doi: 10.3390/app142210294.

[19] S. Das and R. K. Addya, "Skin Cancer Detection and Classification Through Medical Image Analysis Using EfficientNet," NDT, vol. 3, no. 4, article 23, 2025, doi: 10.3390/ndt3040023.