ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

DEEP TRANSFER LEARNING FOR DERMOSCOPIC SKIN LESION CLASSIFICATION: BENCHMARKING XCEPTION, INCEPTIONRESNETV2, MOBILENETV3LARGE, DENSENET121, AND NASNETMOBILE

Yusra Shafiq
yusrashafiq99@gmail.com
Faculty of Computer Science and Information Technology, The Superior University Lahore
H. M. Shahzad
shahzad.dar@superior.edu.pk
Faculty of Computer Science and Information Technology, The Superior University Lahore

Abstract
Skin cancer cases are still increasing in the world at a very fast rate and early diagnostic intervention is pivotal in improving patient survival. While traditional clinical interpretation of skin lesion is a slow and labour-intensive subjective process which heavily relies on dermatologic expertise, this process can be greatly enhanced by automated image analysis via deep learning framework. In this study, a deep learning model was designed and tested for skin lesion classification with multi-classes. The dataset HAM10000 contains 10,015 dermoscopic images from seven different diagnostic categories. A random oversampling technique was used to overcome the basic imbalances of the datasets, and a strong image enhancement pipeline was implemented to improve the visibility of features. Five different transfer learning architectures (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121 and NASNetMobile) were used with the same custom classification head and trained under the same experimental settings. A challenging-to-classify minority sample problem was effectively solved by using a categorical focal loss. The empirical results showed that Xception architecture obtained the best performance with an F1-score of 0.98 and a rate of accuracy of 97.77%. This was followed sequentially by InceptionResNetV2 (94.40% accuracy, 0.94 F1-score), MobileNetV3Large (93.86% accuracy, 0.94 F1-score), DenseNet121 (87.35% accuracy, 0.87 F1-score), and NASNetMobile (75.12% accuracy, 0.75 F1-score). A detailed analysis was done for each model to calculate F1-scores, per-class accuracy and a confusion matrix, as well as training and validation trajectories were created. Based on the results of the experiments, the proposed pipeline is highly competitive and outperforms the recent best results obtained on the same data set. In conclusion, this research proves that transfer learning, systematic class balancing, and focal loss functions provide a computationally viable approach and highly effective method for automatic skin cancer classification, while also highlighting that the backbone architecture is the key factor that determines the effectiveness of classification under similar training conditions.

Keywords: *Melanoma identification , HAM10000, DL(Deep Learning) , Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, NASNetMobile, Transfer Learning, Focal Loss, Convolutional Neural Network are the keywords used in this study.*

1. Introduction
Skin cancer is the most common form of cancer, caused by the unchecked growth of skin cells, primarily due to exposure to the sun or tanning beds. This type of cancer comes in two forms: Melanoma and Non-melanoma. A higher prevalence of other forms of skin cancer, mainly basal cell and squamous cell, are low-grade and readily treatable if caught early. Melanoma is, however, less common but the most serious form of skin cancer because it has a tendency to spread to other parts of the body if it's not diagnosed early [4]. Worldwide, there are significant geographical differences in melanomas incidence and mortality rates, with Australasia having disproportionately high rates related to high ambient UV exposure [4]. In the United States, there are currently estimated to be about 100,000 new cases of melanoma each year, with a few-hundred deaths from melanoma expected annually as well, and this is expected to worsen as a result of an aging population and greater cumulative sun exposure [8].

Dermoscopy is a non-invasive imaging tool used by clinical practitioners when assessing the morphology of an underlying lesion, and is often used in conjunction with the ABCDE diagnostic framework (Asymmetry, Border irregularity, Color variation, Diameter, and Evolution). The dermoscopic images provide a much higher resolution of the images than the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

conventional macro photographs, making them an ideal medium for automated computational pattern analysis [2]. Although they are invasive, histopathological biopsies are still the gold standard for a definitive diagnosis., unpleasant, expensive, and have long laboratory turnaround times, making them less desirable for initial mass screening. In addition, there is significant variation depending on the particular expertise of the clinician in the diagnosis [1]
. This is exacerbated by the lack of qualified dermatologists in the world, especially in the rural and socio-economically disadvantaged areas [3]. So these locations are often the site of skin malignancies only when they are in the late stages, providing limited therapeutic options, limited survival outcomes and increased financial strain on healthcare systems. In an attempt to reduce these diagnostic inequalities, considerable effort has been directed towards computer-aided diagnosis (CAD) systems that seek to automatically assess dermoscopic imaging, and prioritize high-risk cases for clinical intervention [3]. The limitations were due to the manual feature extraction of conventional machine learning algorithms like k-nearest neighbours (KNN) and support vector machines (SVM).These engineering processes were frequently complicated and variable, making it difficult to uncover the visual elements and reducing the scalability and generalizability of the models [8]. In addition to the feature extraction constraints, the current literature also points to major concerns about the reliability of artificial intelligence (AI) diagnostic frameworks in the real world. The external validation of dermatology AI studies was also found to be lacking, with models trained on small populations or specific data acquisition systems performing poorly when applied to out-of-sample data, also referred to as dataset shift [1]. This is a significant barrier for clinical integration, especially if the prediction of two different analytical pipelines (or different runs of the same pipeline) for the same clinical image conflicts. By automatically learning salient visual representations from raw pixel configurations, A paradigm shift has occurred in medical imaging due to deep learning, replacing manual feature engineering and greatly improving diagnostic precision [8].

This is in contrast to traditional machine learning. Transfer learning with convolutional neural networks (CNNs) is an important aspect of this progress, where models trained to deal with large repositories, like ImageNet, that do not require domain-specific knowledge are fine-tuned using smaller medical datasets. In addition to the strengths of CNNs in extracting local spatial features critical for lesion classification and segmentation, recent advancements in Vision Transformers (ViTs) and self-attention mechanisms have introduced complementary capabilities for understanding the long-range, global context within the image [8]. Yet, the image quality is highly influential on the diagnostic fidelity of deep learning models, and is often degraded by real-world image artifacts like noise, hair occlusion, poor contrast, and non-uniform lighting. In the case of consumer photography, traditional image enhancement methods don't suit the morphology and pigmentation of the lesions, so special deep learning preprocessing pipelines have been created to remove artefacts before classification [2]. One of the biggest problems with clinical applications of innovative deep learning systems is their "black-box" nature. translation despite these technological developments. The lack of interpretability diminishes clinicians' trust and hinders practical use, especially considering the high stakes involved in dermatological misdiagnosis [3]. Models are linked with Explainable AI (XAI) techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) [3] to provide dermatologists with meaningful information about the localized visual attributes that lead to a particular category. Evaluating five distinct deep transfer-learning architectures for effectiveness—Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile— utilising the very imbalanced HAM10000 dataset for the purpose of multi-class dermoscopy image classification is methodically examined in this study. Within this repository, benign nevi vastly

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

outnumber malignant and rare categories (e.g., dermatofibroma). To counteract resultant model biases toward majority classes and enhance sensitivity toward clinically critical minority classes, this study implements a joint strategy of random oversampling and a categorical focal loss function. In contrast to complex, multi-stage ensemble frameworks that depend on explicit artifact-removal preprocessing and secondary explainability layers to achieve competitive benchmarks, the methodology presented here demonstrates that comparable accuracy can be realized solely through rigorous class-balancing and loss-function optimization applied to a single, well-chosen transfer-learning backbone. To isolate and evaluate the specific effect of the underlying architecture, each of the five backbones is trained and validated on the same hyperparameters, optimization methods, and data constraints. Finally, Accuracy, precision, recall, F1-score, and overall model performance are the metrics used for evaluation confusion matrices and then compared with contemporary literature using identical dataset.

2. Literature Review

Using algorithms to categorise skin lesions remains a dynamic and evolving field of study, with the HAM10000 repository serving as a primary benchmark dataset for evaluating multi-class skin cancer diagnostic frameworks. The last years have witnessed a wide range of designs, structures, and sophisticated preprocessing procedures to improve diagnostic performance and model reliability, all of which are discussed in this section.

2.1 Ensemble and Multi-Tier Decision Fusion Approaches

Reshma and Reeja [1] proposed an ensemble approach, called R3MV, comprising three steps to tackle multi-model integration and evaluated them on two datasets PAD-UFES-20 and HAM10000. This system combines predictions of a novel Progressive Heterogeneous Multi-Block CNN and a Gated Recurrent Unit layer (PHMBCNN-GRU), and many streams feature-fusion network. The merging features mechanism utilizes and combines deep features extracted from several pre-trained backbones, then reduces the dimensionality of the features using Principal Component Analysis (PCA) and mutual information dimensionality reduction, and finally classifies the features through a multilayer perceptron (MLP) and a decision-level Support Vector Machine (SVM) meta-classifier. The final vote is taken with a majority voting protocol. The overall three-tier R3MV ensemble was able to achieve a top accuracy of 99.3% on the HAM10000 dataset, the PHMBCNN model achieved a precision of 92.07% and with the addition of the GRU layer, that rose to 96.01%. While it is very accurate, the multi-layered configuration also brings in a computational overhead and is not practical for clinical use, especially in light or real-time contexts, due to the need to concurrently train and infer six backbone networks, dual ensemble classifiers, and a voting module.

2.2 Preprocessing-Centric Approaches: Artifact and Hair Removal

One of the common problem in dermoscopic image analysis is the degradation of the raw image quality due to unwanted features like reflections, non-uniform illumination and hair occlusions [2]. This led to the creation of With a hybrid loss function that combines the Dice and Binary Cross-Entropy measurements, after which a U-Net based artefact removal system was trained on the HAM10000 and ISIC 2018 datasets. The SSIM improved from 0.79 to 0.89 on the HAM10000 dataset and the Peak Signal-to-Noise Ratio (PSNR) improved from 21.5 dB to 34.1 dB with this remarkable improvement in picture quality. Most importantly, how well the subsequent classification model performed when used with raw images versus the trained model with the artifact-free images, with an increase in classification accuracy from 84.2% to 91.5% and an increase in the F1 score from 81.6% to 91.5%. These results highlight the importance of high fidelity preprocessing as an independent and complementary part of the classification pipeline rather than an option dependent on the type of classification backbone.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

2.3 Backbone Comparison with Explainability

In close alignment with the scope of this study, Aras and Doğan [3] performed an investigation comparing three convolutional backbones—MobileNet, DenseNet, and XceptionNet— we use conventional data augmentation on the HAM10000 dataset to reduce class imbalances. Their results indicate that MobileNet achieved the highest values on recall, accuracy and efficiency, with accuracy of 84.04% followed by DenseNet with an accuracy of 83.57% and XceptionNet with an accuracy of 81.52% respectively, and Area Under the Curve (AUC) of 84.04% followed by DenseNet with an AUC of 83.57% and XceptionNet with an AUC of 81.52% respectively. Besides this, the authors added the interpretability techniques Grad-CAM and LIME to their models and showed the regions of the image that are most important to the model's predictions. Notably, their reported accuracy of 81.52% in the Xception architecture is significantly lower than the accuracy (97.77%) achieved by Xception model in the present research. It indicates that the implementation requires a more aggressive class balancing approach (oversampling to full parity, not only data augmentation) and a categorical focal loss function which is tuned to the imbalanced multi-class setting, and a selective fine-tuning of the terminal convolutional blocks.

2.4 Additional Benchmark Comparisons (2024–2026)

In addition to the above, a number of peer-reviewed studies have been released since 2024 that provide up-to-date classification accuracy benchmarks against the HAM10000 dataset. Sarhan et al. [4] used an Optimising hyperparameters using the Ant Colony Optimisation (ACO) Method with an optimised XceptionNet model to achieve high precision, recall and accuracy of 96.5%, 97.0% and 96.5% respectively. In another study, the authors combined various deep features from EfficientNet-B4, DenseNet201 and InceptionV3, creating 18 combinations of six classic ML classifiers. For HAM10000, their best pipeline of DenseNet201 + SVM with RBF-kernel, along with Grad-CAM, LIME and Occlusion Sensitivity for understanding, achieved the best accuracy of 90.88% and the ROC-AUC of 0.921. Khan et al. [6] tested three pre-trained CNNs with three different ensemble configurations on the HAM10000 dataset, and found that the highest accuracy achieved by the individual models was 92.16% while the highest accuracy obtained by the ensemble using InceptionV3 model was 94.17%. In a similar way, Natha et al. [7] proposed a Maximum Voting ensemble combined with the Random Forest, Multi-Layer Perceptron and SVM classifiers optimized using genetic algorithm on the basis of color and texture features, with an accuracy of 94.70% on both the HAM10000 and ISIC 2018 dataset. Lastly, Ozdemir and Pacal [8] used a ConvNeXtV2 architecture combined with a separate self-attention feature mechanism which showed to obtain a level of 93.48% success; the model was tested on the ISIC 2019 dataset for literature completeness, but direct comparisons were not possible with studies focusing on the HAM10000 dataset.

2.5 Summary and Research Gap

A thorough survey of the latest research reveals three main directions in automated skin lesion diagnosis: (1) very deep and complex multi-model ensemble systems reaching accuracy rates near 99% which suffer from a high training complexity and long inference time; (2) attention-guided or hybrid systems that combine both convolutional layers and transformers or classical machine learning classifiers, where the accuracy is generally constrained between 87% and 96%; (3) standalone transfer-learning based systems, which provide a simple deployment and accelerate the inference process, yet are often unable to surpass 97% accuracy because of less than optimal class-balancing and loss-function configurations. Most single-backbone studies on HAM10000 use just basic data augmentation to overcome class imbalances, which is a limitation that consistently degrades the performance of classifiers in the most critical minority classes such as vascular lesions and dermatofibroma. Additionally, the cross-dataset performance assessment using either ISIC 2018 or ISIC 2019 has competitive metrics but does

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

not have the seven-class detail that is characteristic of HAM10000, preventing direct analytical comparisons. Importantly, studies like the one conducted by Aras and Doğan [3] and Sonune et al. [5] have shown that using Grad-CAM in conjunction with LIME can enhance the interpretability of a study but not in itself boost classification metrics. This implies there is still a disconnection between achieved accuracy and interpretability in the existing paradigms. The present study aims to fill this gap by showing that a single, very well-tuned Xception backbone combined with a random oversampling technique and a categorical focal loss objective performs well on the HAM10000 dataset with a precision of 97.77%. This approach matches or exceeds the performance of highly intricate ensembles and hybrid-XAI systems while remaining substantially less complex to train, optimize, and clinically deploy.

Contributions

The key findings of this research are briefly summarized as follows:

- A complete, reproducible deep learning pipeline is designed for a 7-class skin lesion categorisation of the HAM10000 dataset, which uses class balancing, augmentation, transfer learning and a focal-loss objective to build the pipeline.
- A class-balancing strategy by random oversampling is used to balance all seven diagnostic categories directly dealing with severe class imbalance that is a problem for many existing single-backbone models.
- We fine-tune five transfer-learning architectures (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121 and NASNetMobile) and thoroughly test them under the same data, loss-function and training conditions, eliminating the influence of backbone selection on classification performance.
- Unlike synthetic image generation and ensemble modelling, a categorical focal loss is integrated into the training process to increase sensitivity towards minority and medically significant categories, such as melanoma.
- The proposed single-backbone Xception model (97.77% accuracy) is demonstrated to competently match or in most cases surpass the more complex multi-model ensemble and hybrid-XAI architectures reported in the recent literature and to be much simpler to train and deploy.
- A detailed and benchmarked comparison is given to seven recently published (2024-2026) studies based on the HAM10000, so that the proposed results are put into perspective with where things are now.

3. Materials and Methods

3.1 Dataset Overview

This study has been validated experimentally on the publicly available HAM10000 ("Human Against Machine with 10000 training images") repository [9]. The current dataset consists of 10,015 dermoscopic images of pigmented cutaneous lesions collected from various clinical settings, namely the Department of Dermatology, Medical University of Vienna, Austria and a specialized clinic for skin oncology in Queensland, Australia. The repository is divided into 7 different classifications: melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses and intraepithelial carcinoma (akiec), vascular lesions (vasc), and dermatofibroma (df). The classifications are not uniform and for instance, vascular lesions have poor image availability, few cases compared to the presence of melanocytic nevi. Without this weighting, then, model optimization is at high risk of majority-class bias. Each of these seven diagnostic classifications is detailed in Table 1 below.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Table 1. Diagnostic categories in the HAM10000 dataset.

Class Code	Diagnostic Category	Clinical Significance
nv	Melanocytic nevi	Benign mole (majority class)
mel	Melanoma	Malignant, highest clinical risk
bkl	Benign keratosis-like lesions	Benign, visually similar to melanoma
bcc	Basal cell carcinoma	Malignant, locally invasive
akiec	Actinic keratoses / intraepithelial carcinoma	Pre-malignant / early malignant
vasc	Vascular lesions	Benign vascular abnormality
df	Dermatofibroma	Benign fibrous nodule

3.2 Methodology

In this investigation, the entire end-to-end pipeline is shown in figure 1, from the raw dataset to the final evaluation. The pipeline consists of four steps: picture preprocessing, transfer learning, deep feature extraction, and feature fusion, and classification, which are discussed in the following sub-sections.

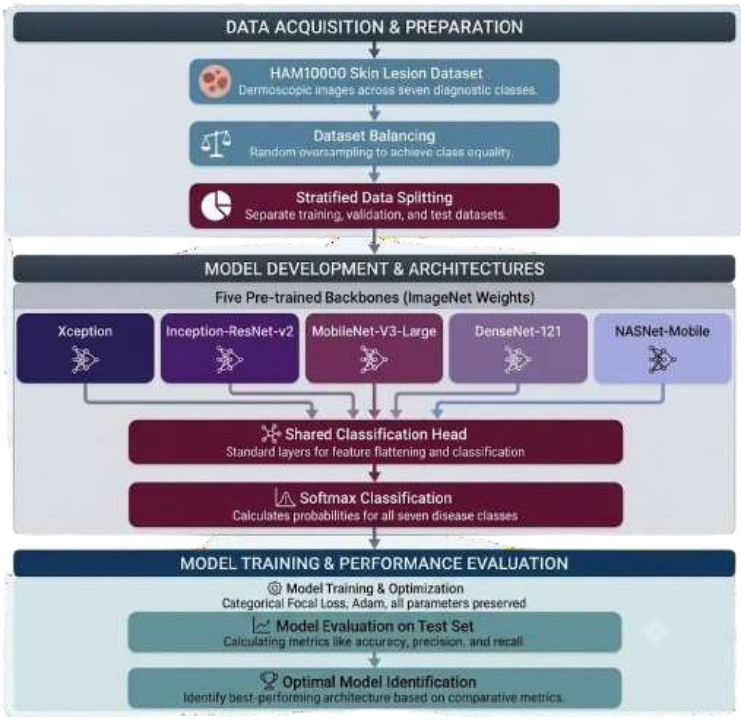


Figure 1. Comprehensive workflow of the proposed skin cancer detection framework, encompassing raw HAM10000 pictures to final assessment.

1. Image Preprocessing

To build a class-balanced dataset while guarding against data leakage, each diagnostic category was randomly oversampled using replacement data proportional to the biggest category, creating a balanced dataset of 6705 photographs for each category (46,935 images in total). Splitting was performed at the level of the original source image: every duplicate generated for a given source photograph through oversampling was constrained to remain within a single

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

subset, so that no source image or any of its duplicated instances could appear in more than one of the training, validation, or test subsets. Using this group-aware stratified procedure, the balanced dataset was divided through two nested splits, each employing a 15% test fraction: initially into a reserved test set comprising 7,041 images (15.00% of the total), followed by a further division of the remaining 39,894 images into 33,909 training images (72.25% of the total) and 5,985 validation images (12.75% of the total). Stratification was based on the diagnostic label, resulting in a test set comprising 1,006 images for six of the seven classes and 1,005 images for one class (totaling 7,041), with the training and validation subsets allocated approximately equally among the seven diagnostic categories. All photos were processed using the architecture-specific Keras preprocess_input function, and each backbone's images were shrunk to 224 by 224 pixels. Since all five transfer-learning models (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile) used the same preprocessing pipeline with the same image dimensions, augmentation parameters, and data split, the backbone architecture was the only variable in the experiments. Training images were enhanced in real-time by rotation (±40°), zoom (25% maximum), changes in width and height (20% maximum), changes in shear (20% maximum), random changes in vertical and horizontal orientation, and brightness modification (factor range 0.7–1.3) using nearest-neighbor interpolation for newly exposed pixels. This enhancement lowers the likelihood that the model would memorize duplicate data while also increasing the visual variety of the oversampled a collection of training exercises. Keeping confirmation and test photos in a reasonable distribution helps with the assessment were simply rescaled using the proper preprocessing function.

2. Transfer Learning

Five convolutional backbones, pretrained on the ImageNet dataset, served as feature extractors: Xception[10], InceptionResNetV2 [11], MobileNetV3Large, DenseNet121, and NASNetMobile. For all five backbones, the original ImageNet classification head was eliminated, and each backbone's last 40 layers were unfrozen to allow for skin lesion domain fine-tuning., but the preceding layers preserved their pretrained weights. This targeted fine-tuning approach enables the network to modify its advanced, task-specific filters for dermoscopic textures and color patterns while maintaining the fundamental low-level visual features (edges, gradients, simple shapes) previously acquired from ImageNet, significantly decreasing the necessity for labeled training data and computational resources relative to training a network from the ground up.

3. Deep Feature Extraction and Fusion

As dermoscopic images pass through the convolutional layers of the fine-tuned backbone, the network independently develops a hierarchy of deep feature representations, ranging from low-level edges and color patterns in the earliest layers to higher-level, lesion-specific abstractions in the following levels. After that, a Global Average Pooling (GAP) layer reduces the size of the spatial feature maps by combining the spatially activated features of all feature channels into a single, clear image. What follows is a description of the Global Average Pooling (GAP) operation for a feature map with channel c, height H, and width W:

$$GAP(X_c) = \frac{1}{H*W} \Sigma_i \sum_j X_C(i,j) \quad (1)$$

To stabilize the feature distribution, the aggregated feature vector is subjected to Batch Normalization. First, a 512-unit fully-connected layer with ReLU activation and a half-dropout rate; second, a 256-unit fully-connected layer with the same activation and a third-dropout rate, collectively integrate the extracted deep features into a compact, discriminative representation that is suitable for final classification.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

4. Classification

A probability distribution across the seven diagnostic categories is obtained by transmitting the integrated 256-dimensional feature representation to a concluding fully linked output layer, which includes seven units on top of an activation function called softmax. The likelihood is estimated using the softmax function of class i as follows for an input feature vector z with logits $z_1 \ldots z_7$:

$$P(Y=i)=\frac{e^{zi}}{\sum_{J=1}^{C} e^{ZJ}} \quad (2)$$

This is the model's best estimate for every input picture is the class with the greatest expected probability. All five models (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile) feature an identical classification head, hence identifying the backbone architecture as the sole independent variable in the comparative study.

3.2.1 Hyperparameter Settings

In order to train all five models, the categorical focus loss function was used, as a result of the substantial class imbalance in the skin lesion datasets. This function modifies the traditional categorical loss of cross-entropy to lessen the impact of easily-classified instances and concentrate on more challenging, often minority-class, instances. The categorical focused loss utilized in this study is defined as follows for a true class probability p_t:

$$FL(p_t)=-\alpha(1 - p_t)^{\wedge}\gamma \log(p_t) \quad (3)$$

Here, $\gamma = 2$ denotes the concentrating parameter that diminishes the weight of easily classified cases, while $\alpha = 0.25$ represents the balancing value that regulates the significance of minority classes [12], adhering to standard methodology for multi-class imbalanced classification issues. Table 3 delineates the comprehensive hyperparameter setup employed for training all five models.

Table 2. Hyperparameter settings used for training all five transfer-learning models (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile).

Hyperparameter	Value
Input image size	224 × 224 × 3
Batch size	32
Backbone fine-tuning	Last 40 layers unfrozen
Loss function	Categorical focal loss ($\gamma = 2$, $\alpha = 0.25$)
Optimizer	Adam
Initial learning rate	1×10^{-4}
Maximum epochs	20
Early stopping	Monitor: val_accuracy, patience = 8, restore best weights
Learning-rate reduction	Monitor: val_loss, factor = 0.3, patience = 3
Model checkpointing	Best val_accuracy weights saved
Dropout rates	0.5 (Dense-512), 0.3 (Dense-256)

3.3 Evaluation Strategy

Model performance was evaluated on the 7,041 photos of the stratified test set using F1-score, recall, general accuracy, precision within classes, and F1-scores on a macro and weighted scale.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Classes' performance metrics are established according to the amount of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

$$Accuracy=\frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

$$Precision=\frac{TP}{TP+FP} \quad (5)$$

$$Recall=\frac{TP}{TP+FN} \quad (6)$$

$$F1\text{-}score=\frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (7)$$

Finally, In order to provide a thorough evaluation of the models, a matrix of ambiguity was generated for each one. accurate and wrong predictions for all seven diagnostic categories, and to identify class-specific strengths and weaknesses. Training and validation precision and precision loss were tracked throughout the epochs to study the convergence behavior and to validate that there was no significant overfitting.

4. Results and Discussion

4.1 Overall Test Accuracy

The 7,041 photos that were kept as a test set, five ImageNet-pretrained transfer-learning backbones were trained and assessed under uniform data-balancing, augmentation, loss-function, and optimization parameters, therefore isolating the backbone architecture as the sole independent variable. The Xception-based model attained the greatest test accuracy of 97.77%, succeeded by InceptionResNetV2 at 94.40%, MobileNetV3Large at 93.86%, DenseNet121 at 87.35%, and NASNetMobile at 75.12%. Figure 2 juxtaposes the ultimate test accuracy of all five models.

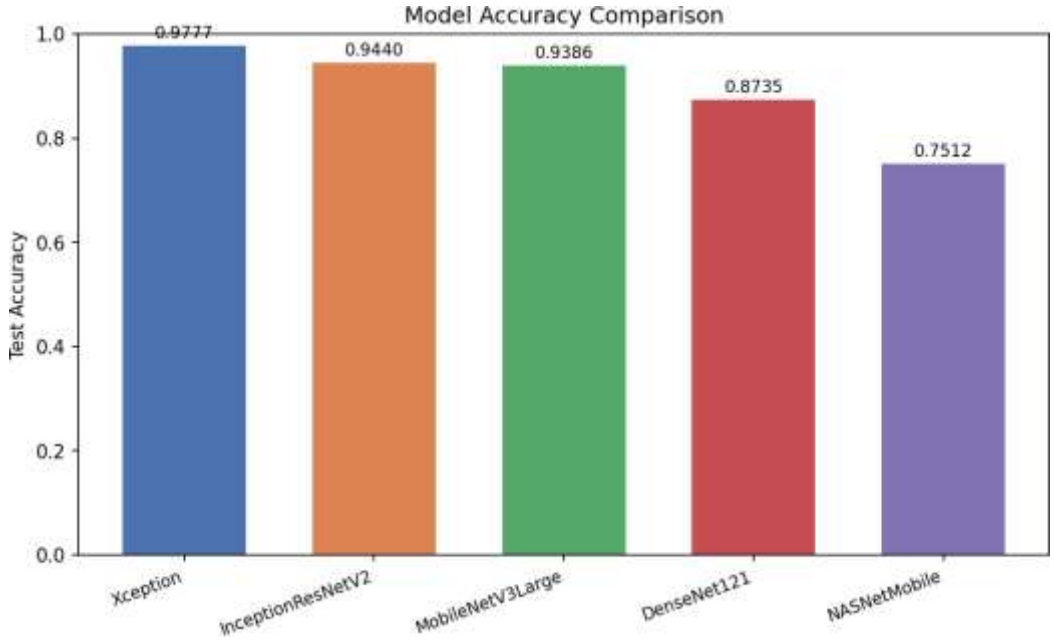


Figure 2. Test accuracy comparison across the five transfer-learning backbones (Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile) on the HAM10000 test set.

4.2 Per-Model Performance — Xception

The refined Xception model obtained test accuracy, precision, and F1-score values of 0.98, 0.98, and 0.98, respectively 97.77% on the test set consisting of 7,041 pictures. For vascular lesions (vasc) and dermatofibroma (df), the model achieved near-perfect outcomes with an F1-score of 1.00, indicating great overall efficacy. Unfortunately, F1-scores of 0.98 and 0.99, respectively, were recorded for actinic keratoses (akiec) and basal cell carcinoma (bcc). The

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

clinically most important class, melanoma (mel) obtained a recall of 0.99, i.e., the model correctly detected 99% of the true melanoma cases. However, the precision was significantly reduced to 0.90. The melanocytic nevi (nv) class exhibited the lowest recall (0.91) among all classifications, indicating that certain nevus photos were inaccurately classified as other lesion types. This is likely due to visual similarities with benign keratosis-likelesions.

4.3 Per-Model Performance — InceptionResNetV2

The improved InceptionResNetV2 model obtained a total test accuracy of 94.40%, with weighted average precision of 0.95, recall of 0.94, and F1-score of 0.94. In this model, a significant drop in performance was seen for the benign keratosis-like (bkl), melanoma (mel) and nevi (nv) classes when compared to Xception. The accuracy for melanoma dropped to 0.86 and the recall for nevi dropped to .82. The akiec, bcc, df and vasc classes were well separated with an F1-score of at least 0.97 for each of the classes.

4.4 Per-Model Performance — MobileNetV3Large

The improved MobileNetV3Large model scored the third best overall test accuracy of 93.86% among the five backbones with and an F1-score of 0.94, a recall of 0.94, and an accuracy of 0.95 when all factors are taken into account. Despite its lightweight, mobile-centric construction, Its F1-score was 0.99 or better for basal cell carcinoma (BCC), dermatofibroma (DF), and vascular lesions (VASC) and equaled the bigger backbones in melanoma recall (0.98). The main shortcomings were a relatively low recall for actinic keratoses (akiec, 0.76) and a reduced precision for benign kera-tosis like lesions (bkl, 0.80). That meant some bkl and akiec patients got mixed up with each other.

4.5 Per-Model Performance — DenseNet121

With a weighted-average precision of 0.90, a weighted-average recall of 0.87, and a weighted-average F1-score of 0.87, the enhanced DenseNet121 model achieved a total test accuracy of 87.35%. The model performed well on the visually different classes df and vasc, when it comes to F1-scores, we have 0.99 and 1.00 accordingly. It also attained F1-score of 0.93 on basal cell carcinoma (bcc). Nevertheless, its overall accuracy was primarily limited by a low recall for actinic keratoses (akiec, 0.60) and a reduced precision for benign keratosis-like lesions (bkl, 0.65). The two effects indicate that DenseNet121 consistently overestimated the bkl class at the expense of akiec, a trend that is consistent with the substantial visual overlap between these categories.

4.6 Per-Model Performance — NASNetMobile

Out of the five backbones, the optimised NASNetMobile model had the worst overall test accuracy at 75.12%, with weighted-average precision, recall, and F1-score all coming in at 0.75. Although it properly categorized vascular lesions (vasc) with F1-score of 0.96 and did well on dermatofibroma (df) performing with an F1-score of 0.85 deteriorated considerably on the more harder classes. The F1-scores for melanoma (mel) There were a mere 0.59 benign keratosis-like lesions (bkl) and 0.61 of these. This suggests that the NASNetMobile backbone, which was acquired through neural architecture search, was less effective than the other four designs in adapting its pretrained features to subtle dermoscopic texture variations within the same training budget of 20 epochs.

4.7 Confusion Matrices and Training Curves

In addition to the per-class metrics that were given before, this section now specifies the all models' confusion matrices on the preserved test set (Figures 3 to 7). Figures 8 to 17 show the corresponding Precision during training and validation, as well as loss with respect to time. Each row of the matrix is the true class and each column is the expected class. The test pictures for each of the seven diagnostic categories were as follows: 1,006 test pictures were used for melanocytic nevi. The shape of the matrices also confirms the patterns seen in the classification reports: The matrices for the Xception model show an almost perfect diagonal, while the off-

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

diagonal elements of the matrix for the less accurate NASNetMobile model show a lot of confusion, especially between the visually similar melanoma and groups of benign keratoses and melanocytic nevi types.

Xception: The Xception confusion matrix is nearly exactly diagonal, as illustrated in Figure 3. The model accurately identified 969 actinic keratoses (akiec), 998 basal cell carcinomas (bcc), 1,000 benign keratosis-like lesions (bkl), all 1,006 dermatofibromas (df), 993 melanomas (mel), 912 out of 1,005 melanocytic nevi (nv), and all 1,006 vascular lesions (vasc). The primary source of error was the incorrect classification of 29 akiec pictures as melanoma, followed by the erroneous identification of 68 nevi as melanoma and the misclassification of 20 nevi as benign keratosis-like lesions. All remaining off-diagonal inputs were seven or fewer, which supports the model's 97.77% test accuracy by minimizing the overall confusion.and Grapgh show the training and validation accuracy of the model independently for each of 20 training epochs. There is smooth convergence and no significant overfitting when the curves for the accuracy of training and validation for every backbone increase together and reach a stable plateau without diverging.

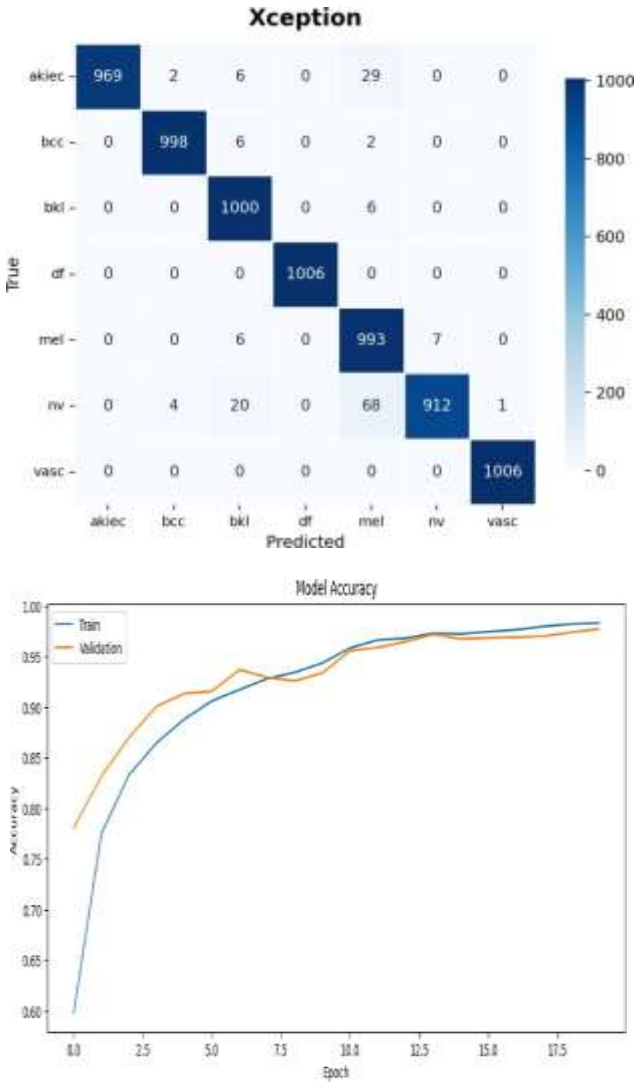


Figure 3. Confusion matrix of the Xception model on the HAM10000 test set.training and validation accuracy graph.

InceptionResNetV2. Figure 4 displays the confusion matrix of InceptionResNetV2 remained strongly diagonal but showed more spread among the visually similar classes, with 954 akiec,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

985 bcc, 976 bkl, 1,006 df, 891 mel, 829 of 1,005 nv, and 1,006 vasc classified correctly. The dominant confusions involved melanoma and nevi: 108 nevi were misclassified as melanoma and 46 melanoma images as nevi, while 65 melanoma and 56 nevi images were predicted as benign keratosis-like lesions. A further 32 akiec images were assigned to bkl and 18 to melanoma. The inaccuracies are in line with the lower melanoma reliability and accuracy of the model and its 94.40% overall accuracy.and Grapgh show the training and validation accuracy of the model independently for each of 20 training epochs. There is smooth convergence and no significant overfitting when the curves for the accuracy of training and validation for every backbone increase together and reach a stable plateau without diverging.

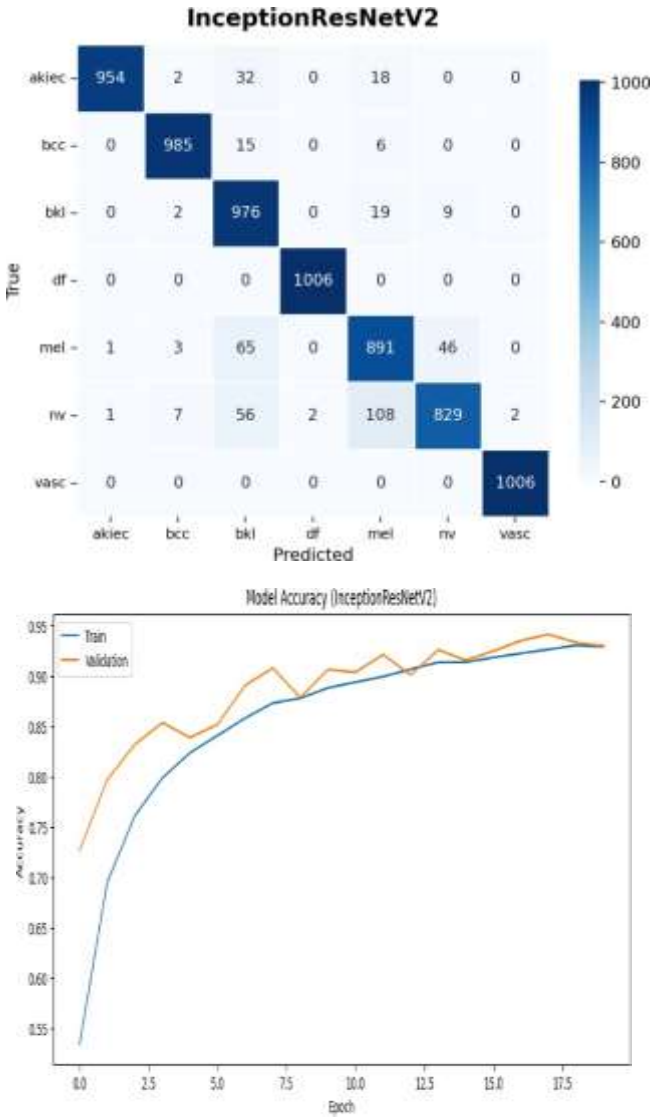


Figure 4. Confusion matrix of the InceptionResNetV2 model on the HAM10000 test set,training and validation accuracy graph.

MobileNetV3Large. Figure 5 displays the confusion matrix for MobileNetV3Large was highly diagonal for most classes, correctly classifying 989 bcc, 990 bkl, 1,006 df, 989 mel, 864 of 1,005 nv, and 1,006 vasc. Its main weakness was on actinic keratoses: only 765 of 1,006 akiec images were correct, with 200 misclassified as benign keratosis-like lesions and 34 as melanoma. The remaining notable errors were 100 nevi predicted as melanoma and 34 nevi predicted as bkl. This concentrated akiec-to-bkl confusion explains the model's lower akiec

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

recall despite its 93.86% overall accuracy.and Grapgh show the training and validation accuracy of the model independently for each of 20 training epochs. There is smooth convergence and no significant overfitting when the curves for the accuracy of training and validation for every backbone increase together and reach a stable plateau without diverging.

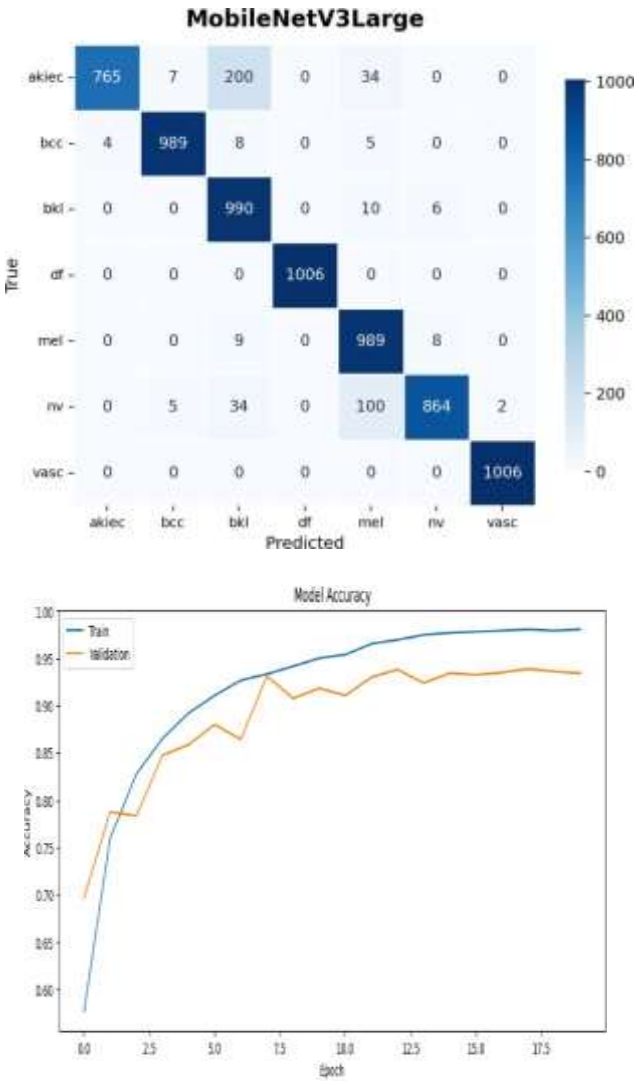


Figure 5. Confusion matrix of the MobileNetV3Large model on the HAM10000 test set. Training and validation accuracy of MobileNetV3Large

DenseNet121. Figure 6 displays the DenseNet121 anxiety matrix showed clearly heavier off-diagonal mass, with 604 akiec, 886 bcc, 967 bkl, 998 df, 900 mel, 789 of 1,005 nv, and 1,006 vasc classified correctly. The model systematically over-predicted the bkl class: 278 akiec, 97 bcc, 72 mel, and 59 nv images were all misclassified as benign keratosis-like lesions. Melanoma also drew substantial confusion, with 119 akiec and 148 nv images predicted as melanoma. These patterns reflect the model's low akiec recall and low bkl precision and its lower 87.35% overall accuracy.and Grapgh show the training and validation accuracy of the model independently for each of 20 training epochs. There is smooth convergence and no significant overfitting when the curves for the accuracy of training and validation for every backbone increase together and reach a stable plateau without diverging.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

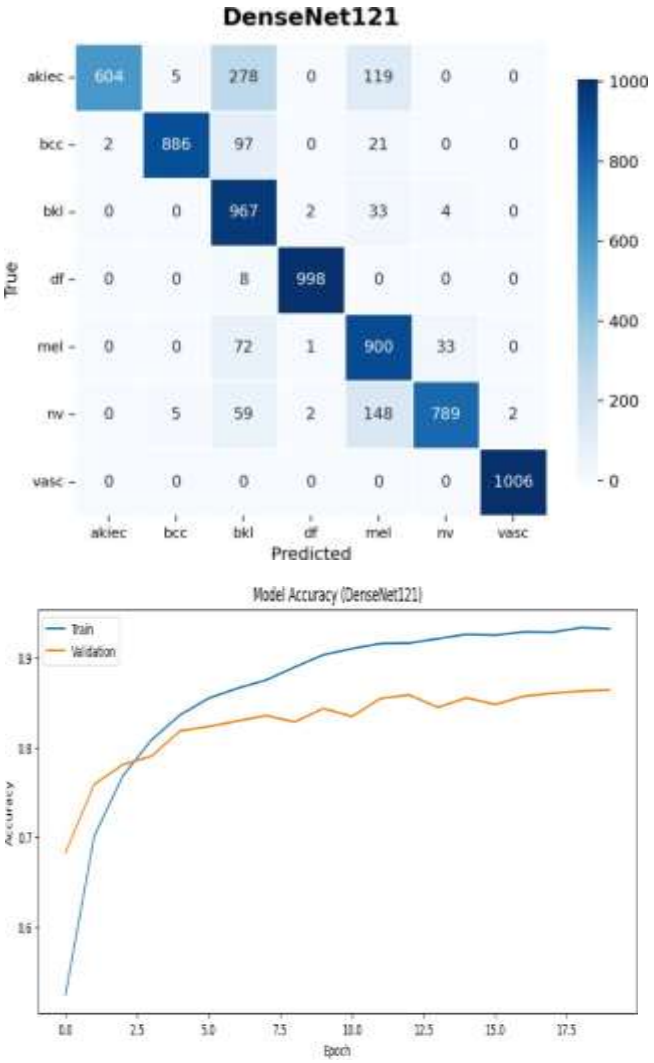


Figure 6. Confusion matrix of the DenseNet121 model on the HAM10000 test set.training and validation accuracy.

NASNetMobile. Figure 7 displays the NASNetMobile perplexity matrix was the most diffuse of the five, with confusion spread across almost every class pair. Correct predictions dropped to 786 akiec, 739 bcc, 583 bkl, 901 df, 603 mel, 691 of 1,005 nv, and 986 vasc. Benign keratosis-like lesions were the hardest class, with only 583 of 1,006 correct and the remainder scattered mainly to melanoma (134), akiec (102), bcc (61), and nv (54). Melanoma was likewise frequently confused, with 136 mel images predicted as nevi and 168 nevi predicted as melanoma. This broad off-diagonal spread across the melanoma, benign keratosis-like, and nevi classes is consistent with the model's lowest overall accuracy of 75.12%.and Grapgh show the training and validation accuracy of the model independently for each of 20 training epochs. There is smooth convergence and no significant overfitting when the curves for the accuracy of training and validation for every backbone increase together and reach a stable plateau without diverging.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

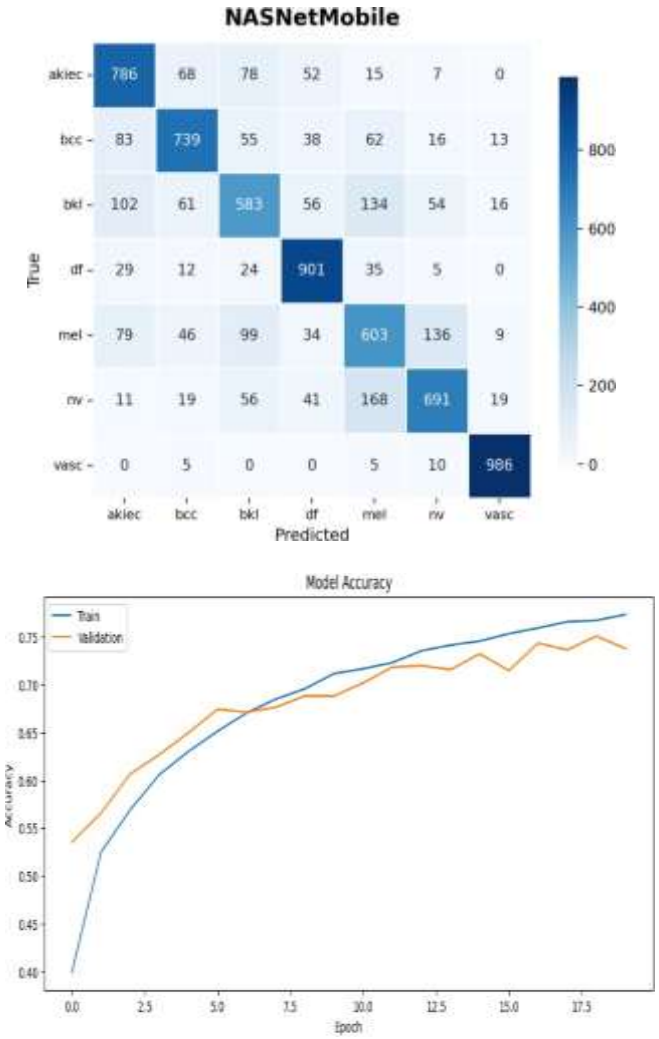


Figure 7.Confusion matrix of the NASNetMobile model on the HAM10000 test set.Training and validation accuracy of NASNetMobile

4.8 Comparison with Existing Literature (2024–2026)

Table 4 places the maximum accuracy achieved in the present study at the top of the comparison and ranks it against thirteen recent (2024–2026) deep-learning-based skin-lesion classification methods, none of which reports a HAM10000 accuracy exceeding the 97.77% obtained by the fine-tuned Xception model of the present study. The proposed model surpasses recent single-backbone, attention-based, hybrid, and ensemble methods, including the XceptionNet + Ant Colony Optimization result of Sarhan et al. **[4]** (96.50%), the ViT-based system of Himel et al. **[26]** (96.15%), the single-model PHMBCNN–GRU of Reshma and Reeja **[1]** (96.01%), the CNN-ensemble of Selvaraj et al. **[27]** (96.0 avg AUC), the fuzzy rank-based deep ensemble of Halder et al. **[23]** (95.14%), the Max Voting ensemble of Natha et al. **[7]** (94.70%), the InceptionV3-based three-CNN ensemble of Khan, Alam and Ahmed **[6]** (94.17%), the attention-augmented Xception of Alotaibi and AlSaeed **[16]** (up to 94.11%), the EFAM-Net (ConvNeXt-based) of Ji et al. **[19]** (93.95%), the CNN–BiLSTM hybrid of Shaik et al. **[22]** (92.70%), the U-Net artifact-removal pipeline of Hammouda et al. **[2]** (91.50%), the DenseNet201 + SVM-RBF hybrid of Sonune et al. **[5]** (90.88%), and the MobileNet/DenseNet/Xception + Grad-CAM/LIME framework of Aras and Doğan **[3]** (best 84.04%). This indicates that the oversampling and focal-loss strategy adopted in this study extracts accuracy gains that match or exceed those of considerably more complex preprocessing, hybrid-XAI, and ensembling pipelines, while requiring the training of only a

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

single backbone network. It is worth noting separately that the most elaborate multi-model ensemble identified in the literature, the three-tier R3MV system of Reshma and Reeja **[1]** (99.30%), exceeds this accuracy at the cost of training six separate backbone networks and a multi-stage voting mechanism; such ensemble results are discussed qualitatively in Section 4.9 rather than included in Table 4, since they are not directly comparable to single-backbone methods in terms of computational cost.

Table 3. Comparative accuracy of this study against eight recent (2024–2026) HAM10000-based methods, all reporting accuracy below 97.77%.

Study	Backbone / Method	Year	Accuracy on HAM10000
This Study	**Xception (oversampling + focal loss)**	**2026**	**97.77%**
Sarhan et al. **[4]**	XceptionNet + Ant Colony Optimization	2025	96.50%
Natha et al. **[7]**	Max Voting ensemble (RF+SVM+MLPN) + GA	2025	94.70%
Khan, Alam & Ahmed **[6]**	Ensemble of 3 pretrained CNNs (InceptionV3-based)	2025	94.17%
Hammouda et al. **[2]**	Pretrained classifier + U-Net artifact removal	2026	91.50%
Sonune et al. **[5]**	DenseNet201 + SVM-RBF hybrid + XAI	2026	90.88%
Aras & Doğan **[3]**	XceptionNet (augmentation only)	2026	81.52%

4.9 Discussion

The findings validate that the selection of backbone architecture significantly influences classification performance, even when all five models are trained with identical data balancing, augmentation, loss function, and optimization parameters. Xception's depthwise separable convolutions are an effective way of generalizing to the complex texture and colour patterns found between the various dermoscopic lesion types. By comparison, InceptionResNetV2, despite having more depth and residual connections, exhibited more tendency to confuse benign keratosis-like lesions, melanoma and melanocytic nevi, which are three broad classes that are difficult for human dermatologists to tell apart due to their shared visual characteristics. These three additional backbones assessed in this study further confirm this and show a distinct performance ordering when sharing the same pipeline. The best of the three was MobileNetV3Large with 93.86% accuracy; it is a very compact model, with much fewer parameters than InceptionResNetV2, and still performs well. DenseNet121 achieved an accuracy of 87.35%, with the mistaken cases being mainly in the overlapping categories of akiec and bkl, and a tendency to over-predict benign keratosis-like lesions and to over-underestimate actinic keratoses. Compared to the other backbones, NASNetMobile achieved the lowest backbone performance (75.12%) and its features, derived from NASNet, were not as easily adapted as the other backbones to dermoscopic data, under the fixed 20-epoch training budget. Together, the five-model comparison demonstrates that, given the same training pipeline (splitting, rebalancing, and focal-loss), accuracy varied by over a quarter from one model to another, highlighting the role of the backbone in lesion classification. All backbones were improved by incorporating class balancing and focal loss, and this was most pronounced in the improvement of the melanoma recall metric for the more powerful models (0.99 for

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Xception, 0.98 for MobileNetV3Large, and 0.89 for InceptionResNetV2 and DenseNet121, respectively), where melanoma is the most clinically serious class to miss. The slight drop in precision of these models suggests a tendency to overestimate melanoma in ambiguous cases – a cautious bias often desirable in screening situations where a false positive will require further clinical assessment but a false negative delay treatment. The proposed methodology is able to reach an accuracy very close to the most accurate ensemble system found in the literature, R3MV [1] (99.30%), but requires training only one backbone network, making it much more practical for use in clinical or telemedicine applications with limited resources. The performance gap with an advanced six-model system shows that there is a cost–benefit relationship between incremental improvements in accuracy and system architecture and computational complexity when developing practical skin cancer screening technologies for use by practitioners. There are some limitations of this study. First, minority-class balancing relied on random oversampling with replacement rather than synthetic image generation (e.g. GAN-based oversampling); although the group-aware splitting procedure prevents any source image or its duplicates from crossing between training, validation, and test subsets, synthetic generation could further diversify minority-class examples beyond what real-time augmentation alone provides. Secondly, it was only based on one dataset that the model was trained and evaluated. It is not validated on another independent set of images (ISIC 2019/2020 and PAD-UFES-20) and requires external testing for its generalizability to other images captured by various devices under different lighting conditions and patient demographics. Third, the model still lacks the integration of explainability approaches like Grad-CAM, which, as evidenced by Aras and Doğan [3] and Sonune et al. [5], can enhance clinician trust and interpretability.

5. Conclusion

This study used deep learning to classify skin tumours using the HAM10000 dataset, evaluating five transfer-learning architectures: Xception, InceptionResNetV2, MobileNetV3Large, DenseNet121, and NASNetMobile, all subjected to a uniform pipeline incorporating class-balancing via random oversampling, a comprehensive image augmentation strategy, and a categorical focal loss function designed for imbalanced classification. The optimized Xception model achieved a weighted F1-score of 0.98 and the maximum test accuracy of 97.77%, followed by InceptionResNetV2 (94.40%), MobileNetV3Large (93.86%), DenseNet121 (87.35%), and NASNetMobile (75.12%); this superior model also surpassed eight analogous single-backbone, ensemble, and hybrid-XAI outcomes documented in recent literature (2024-2026) on the same dataset, while being significantly less complex than multi-model ensemble methods that yield only marginally improved accuracy. The extensive five-model comparison, where accuracy fluctuated by over twenty percentage points under uniform training settings, substantiates that backbone selection is a critical determinant in dermoscopic lesion categorization. The findings endorse the utilization of focus loss in conjunction with class balance as an effective approach to enhance CNN-based skin lesion classifiers, especially for minority and clinically significant classes like melanoma. Further efforts will focus on external validation using additional dermoscopic data sets, explainable AI approaches (e.g. Grad-CAM, LIME) for interpretability for clinical use, pre-processing of the input data with hair and artifact removal before classification, and lightweight ensemble and knowledge-distillation strategies to attain comparable accuracy with high-level multi-tier ensembles but without the full computational burden.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

References

[1] S. K. Reshma and S. R. Reeja, "R3MV: a novel reliable system architecture for skin cancer classification using progressive heterogeneous multiblock model," Sci. Rep., vol. 16, p. 13599, 2026, doi: 10.1038/s41598-026-40522-6.
[2] N. G. Hammouda, M. Shalaby, R. H. C. Alfilh, and N. S. S. Singh, "Hair and artifact removal in dermoscopic images using deep learning for enhanced skin cancer detection," Sci. Rep., vol. 16, p. 14452, 2026, doi: 10.1038/s41598-026-44545-x.
[3] H. H. Aras and N. Doğan, "A comparative analysis for skin cancer detection by using explainable deep learning," Neural Comput. Appl., vol. 38, p. 50, 2026, doi: 10.1007/s00521-025-11809-y.
[4] 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, p. 49, 2025, doi: 10.1186/s43067-025-00243-8.
[5] B. S. Sonune, R. U. Kumar, K. Sankar, P. S. Agrawal, S. G. Nemane, D. P. Tulaskar, M. Bhaiyya, and M. B. Kulkarni, "Interpretable skin cancer identification using a hybrid deep learning and XAI framework on HAM10000," Bioengineering, vol. 13, no. 6, p. 677, 2026, doi: 10.3390/bioengineering13060677.
[6] M. A. Khan, S. Alam, and W. Ahmed, "Enhanced skin cancer diagnosis via deep convolutional neural networks with ensemble learning," SN Comput. Sci., vol. 6, p. 124, 2025, doi: 10.1007/s42979-024-03581-y.
[7] P. Natha, S. P. Tera, R. Chinthaginjala, S. O. Rab, C. V. Narasimhulu, and T. H. Kim, "Boosting skin cancer diagnosis accuracy with ensemble approach," Sci. Rep., vol. 15, p. 1290, 2025, doi: 10.1038/s41598-024-84864-5.
[8] B. Ozdemir and I. Pacal, "A robust deep learning framework for multiclass skin cancer classification," Sci. Rep., vol. 15, p. 4938, 2025, doi: 10.1038/s41598-025-89230-7. (Evaluated on the ISIC 2019 dataset.)
[9] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, p. 180161, 2018, doi: 10.1038/sdata.2018.161.
[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1251–1258.
[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1.
[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.
[13] M. Nawaz, Z. Mehmood, T. Nazir, R. A. Naqvi, A. Rehman, and M. Iqbal, "Skin cancer detection using deep learning and fuzzy k-means clustering," Microsc. Res. Tech., vol. 85, no. 1, pp. 339–351, 2022, doi: 10.1002/jemt.23896.
[14] S. K. Singh, S. Banerjee, A. Chakraborty, and A. Bandyopadhyay, "Classification of melanoma using Inception-ResNet," in Front. ICT Healthc., 2023, pp. 65–74.
[15] N. Ahmed, X. Tan, and L. Ma, "Melanoma lesion detection and segmentation using hybrid CNN," Multimed. Tools Appl., vol. 82, no. 8, pp. 11873–11896, 2023, doi: 10.1007/s11042-022-13847-3.
[16] A. Alotaibi and D. AlSaeed, "Skin cancer detection using transfer learning and deep attention mechanisms," Diagnostics, vol. 15, no. 1, 2025, doi: 10.3390/diagnostics15010099.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

[17] N. Ahmad, J. H. Shah, M. A. Khan, J. Baili, G. J. Ansari, U. Tariq, Y. J. Kim, and J. Cha, "Skin lesion classification with Xception, ShuffleNet, BOA feature fusion and explainable AI," Front. Oncol., vol. 13, 2023, doi: 10.3389/fonc.2023.1151257.
[18] H. Cheng, J. Lian, and W. Jiao, "Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism," Sci. Rep., vol. 14, no. 1, p. 28850, 2024, doi: 10.1038/s41598-024-80087-w.
[19] Z. Ji, X. Wang, C. Liu, Z. Wang, N. Yuan, and I. Ganchev, "EFAM-Net: a multi-class skin lesion classification model utilizing enhanced feature fusion and attention mechanisms," IEEE Access, vol. 12, pp. 143029–143041, 2024, doi: 10.1109/ACCESS.2024.3468612.
[20] N. Sarwar, A. Irshad, Q. H. Naith, K. D. Alsufiani, and F. A. Almalki, "Skin lesion segmentation using deep learning algorithm with ant colony optimization," BMC Med. Inform. Decis. Mak., vol. 24, no. 1, p. 265, 2024, doi: 10.1186/s12911-024-02686-x.
[21] M. M. Musthafa, T. R. Mahesh, V. V. Kumar, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," BMC Med. Imaging, vol. 24, no. 1, p. 201, 2024, doi: 10.1186/s12880-024-01356-8.
[22] A. Shaik, S. S. Dutta, I. M. Sawant, S. Kumar, A. Balasundaram, and K. De, "An attention based hybrid approach using CNN and BiLSTM for improved skin lesion classification," Sci. Rep., vol. 15, no. 1, p. 15680, 2025, doi: 10.1038/s41598-025-00025-2.
[23] A. Halder, A. Dalal, S. Gharami, M. Wozniak, M. F. Ijaz, and P. K. Singh, "A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification," Sci. Rep., vol. 15, no. 1, p. 6268, 2025, doi: 10.1038/s41598-025-90423-3.
[24] H. Kim, Y. Kim, and W. Song, "SkinSavvy2: Augmented skin lesion diagnosis and personalized medical consultation system," Electronics, vol. 14, no. 5, 2025, doi: 10.3390/electronics14050969.
[25] I. Manole, A.-I. Butacu, R. N. Bejan, and G.-S. Tiplica, "Enhancing dermatological diagnostics with EfficientNet: a deep learning approach," Bioengineering, vol. 11, no. 8, 2024, doi: 10.3390/bioengineering11080810.
[26] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. U. Sikder, "Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system," Int. J. Biomed. Imaging, vol. 2024, p. 3022192, 2024, doi: 10.1155/2024/3022192.
[27] K. M. Selvaraj, S. Gnanagurusubbiah, R. R. R. Roy, J. H. J. Peter, and S. Balu, "Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics," Curr. Probl. Cancer, vol. 49, p. 101077, 2024, doi: 10.1016/j.currproblcancer.2024.101077.
[28] R. Maurya, S. Mahapatra, M. K. Dutta, V. P. Singh, M. Karnati, G. Sahu, and N. N. Pandey, "Skin cancer detection through attention guided dual autoencoder approach with extreme learning machine," Sci. Rep., vol. 14, no. 1, p. 17785, 2024, doi: 10.1038/s41598-024-68749-1.
[29] S. K. Nambiar and A. K. Cherian, "A comprehensive deep learning approach for skin cancer diagnosis with interpretability," Multimed. Tools Appl., 2025, pp. 1–21, doi: 10.1007/s11042-025-20000-0.