Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

A LEAKAGE-FREE TWO-PHASE TRANSFER LEARNING ENSEMBLE FOR BINARY MELANOMA CLASSIFICATION USING EFFICIENTNETB3, DENSENET121, INCEPTIONV3, AND VIT-B16

Umair Ayaz Kamangar[1], Abdul Sattar Chan[*2], Zainab Umair Kamangar[3]

[1]*Department of Computer System Engineering, Sukkur IBA University, Sindh, Pakistan*
[*2]*Department of Computer System Engineering, Sukkur IBA University, Sindh, Pakistan*
[3]*Department of Computer Science, Sukkur IBA University, Sindh, Pakistan*

[1]umair.ayaz@iba-sik.edu.pk, [*2]abdul.sattar@iba-suk.edu.pk, [3]zainabumair.phdcss22@iba-suk.edu.pk

DOI: https://doi.org/10.5281/zenodo.20082910

Keywords
Melanoma Detection, Skin Cancer Classification, Transfer Learning, Vision Transformer, EfficientNetB3, DenseNet121, InceptionV3, Ensemble Learning, Two-Phase Fine-Tuning, Deep Learning.

Article History
Received: 14 March 2026
Accepted: 24 April 2026
Published: 08 May 2026

Corresponding Author: *
Abdul Sattar Chan

Abstract
Melanoma, a type of skin cancer, is among the fastest-growing and most lethal cancers worldwide. Automation-based early detection of this disease is very crucial for increasing the survival rates of patients. This paper conducts a thorough comparative study of four pre-trained deep learning architectures: EfficientNetB3, DenseNet121, InceptionV3, and Vision Transformer (ViT-B16) with the aid of a weighted ensemble method for binary classification of melanoma on the Kaggle Melanoma Skin Cancer Dataset containing 10,000 dermoscopic images of 5,000 benign and 4,538 malignant skin lesions. Our work features a well-defined two-stage transfer learning methodology that effectively allows for the prevention of data augmentation leakage by performing stratified splitting before augmentation and significantly improves feature adaptation by resorting to progressive fine-tuning and adaptive learning rate scheduling. Individual model accuracies achieved are: EfficientNetB3 (94.90%), DenseNet121 (95.26%), InceptionV3 (95.11%), and ViT-B16 (96.04%). The weighted ensemble that combines the four models achieves 96.77% accuracy, 0.9886 precision, 0.9656 F1-score, and 0.9949 AUC, exceeding the 95.25% ensemble baseline of Sarıateş and Özbay by 1.52 percentage points on the same dataset. This implies that an effectively designed pipeline can continuously increase the accuracy irrespective of the model architecture, and that an ensemble of complementary CNN and Transformer features leads to even better results than single models.

1. Introduction

Skin cancer ranks among the most prevalent and rapidly increasing types of cancers worldwide. If undetected, its deadliest form, melanoma, will metastasize rapidly [1]. Survival rates at 10 years after diagnosis vary between 93% for early-stage melanoma and as low as 36% for late-stage melanoma [2], which is why a fast and accurate diagnosis remains the major determinant of survival.

While dermoscopic examination along with clinical evaluation and histological biopsy remain the conventional method of diagnosing, they not only demand a lot of time but are also highly variable between observers and depend heavily on the availability of expert knowledge [3]. As a result of these limitations, the development of

automated AI-based systems for the classification of skin lesions that are fast, dependable, and accurate is gaining a lot of attention.

Deep learning methods especially Convolutional Neural Networks (CNNs) have shown outstanding application success in medical image analysis. Networks with pre-trained CNN architectures, fine-tuned using transfer learning, have reached or exceeded the level of expert human diagnosticians in several recent comparative studies [4, 5], and at the recent ISIC 2018 challenge, using ensemble models. More recently, Vision Transformers, which utilize multi-head self-attention over a series of image patches, have shown a similar ability to CNNs to learn relevant but non-localized features that local convolutional filters might not be able to catch, and ensemble models have shown further performance increases through predictive combination [6, 7].

However, there remain several recurring limitations in the current literature on deep learning for melanoma detection: augmented images are duplicated across training and validation fold resulting in data leakage and artificial inflation of accuracy scores [8]; per-model single phase training prevents adaptation to specific domains; CNN-Transformer comparison studies do not abide by a single, consistent experiment pipeline so comparisons are not easily made between studies.

In order to overcome these limitations, this work first applies the systems two-phase transfer learning pipeline in the same manner to the four architectures (EfficientNet B3, DenseNet 121, Inception V3, ViT-B16) and then combines them by a weighted ensemble. The major contributions of this paper are:

1. A data preparation protocol without leakage: original image paths are divided before any augmentation, so that the validation set consists only of unaltered images.
2. A two-stage progressive fine-tuning approach: Stage 1 trains the classifier head with the base frozen; Stage 2 unfreezes the top layers and fine-tunes at a very low learning rate.
3. An adaptive learning rate scheduling method by means of ReduceLROnPlateau for better convergence.
4. A comprehensive comparison of four architectures including a Vision Transformer, demonstrating pipeline methodology accounts for major accuracy differences between studies.
5. A weighted ensemble combining CNN and Transformer predictions, achieving 96.77% accuracy and surpassing the Sarıateş and Özbay [24] ensemble baseline of 95.25% by 1.52 percentage points.

2. Literature Review

Rapid advances in neural network structures, availability of large-scale dermoscopic image datasets, have led to rapid growth in applying deep learning methods to the automated diagnosis of skin cancer.

2.1 Transfer Learning and CNN Architectures

Transfer learning has become the de facto approach to medical image classification, where there is a scarcity of labelled training data. More formally, given a source domain D_S and a learned feature representation ϕ_S, transfer learning adapts the model to a target domain D_T by minimising the target loss;

$$L_T(\phi_S(x);\theta_T), \quad [1]$$

where ϕ_T denotes the task-specific parameters fine-tuned on the target dataset [9].his CNN was pre-trained on ImageNet and modified for specific dermatoscopic tasks by re-training certain layers. Since all the CNN models rely on features pre-learned on ImageNet, they only require training on the specific domain data, significantly reducing the time for training [9]. Renu et al. compared EfficientNetV2, Inception V3 and a generic CNN on the ISIC data set, with EfficientNet V2 achieving an accuracy of 84% [10]. Harahap et al. evaluated the performance of EfficientNet models B0B7 on the ISIC 2019 data set, with the B4 model achieving an accuracy of 75.66% [11]. Using a deep transfer learning approach on Modified EfficientNet B3, Prasad et al. achieved a validation accuracy of 90.6% on a Kaggle based skin cancer data set [12]. Sabir and

Mehmood compared several CNN architectures for melanoma classification on image data, reporting accuracy and loss profiles for various CNN models [13]. Almufareh et al. devised and tested a fine-tuned CNN to identify and classify melanoma, with the layer adaptation of pre-trained images providing consistent performance enhancements [14]. Ahmed et al. reported accuracy of a stacked CNN for melanoma prediction, emphasizing the models ability of extracting convolutional features using several convolutional layers before classifying [15]. Gupta and Mesram put forward a hybrid AlexNet and DenseNet-121 for dermoscopic image based skin cancer prediction with accuracy comparable to other models that leverage deep features [16].

Neeshma and Nair applied DenseNet121 to multiclass skin lesion classification using HAM10000 data, 80.5% accuracy was obtained with an imbalanced dataset and 82.1% accuracy after data resampling [17]. Siddique et al. created a deep sequential CNN for skin cancer detection on HAM10000 training data and attained 96.25% accuracy, outperforming InceptionV3 and ResNet-50 [18]. Hamim et al. created an XAI system called SmartSkin-XAI which added the Grad-CAM XAI to DenseNet121 which resulted in 95% accuracy with the base DenseNet121 and 98% accuracy with the full SmartSkin-XAI system trained on the Kaggle melanoma data set. This system used pre-training from multi-dataset training on the ISIC 2020 data set before fine-tuning on the Kaggle melanoma data set [19].

2.2 Vision Transformers in Medical Imaging

Vision Transformers (ViT), in this approach Dosovitskiy et al. treat an image as a series of fixed-size patches and provide multi-head self-attention to the patch embeddings [6]. ViT provides global context modelling which is not possible with CNNs. Pacal et al. designed a CNN-ViT hybrid which achieved 92.54% accuracy on ISIC 2019 [20].

Toure et al. proposed a hybrid ResNet-ViT model, which reached 95.1% accuracy and 0.971 AUC in melanoma classification [21]. Dagnaw et al. analyzed Vision Transformer architectures with XAI applied in skin cancer, which produced a body of knowledge on how attention maps handle the interpretability task [22]. Kanadath et al. proposed CViTS-Net, a CNN-ViT network with skip connections applied to histopathology classification, which was both performant and efficient [23]. Sarıateş and Özbay, which was the primary baseline of this work, compared DenseNet121, InceptionV3, ViT, and Xception architectures on the Kaggle melanoma dataset. Individual accuracies: DenseNet121 (94.50%), Xception (93.80%), InceptionV3 (91.20%), ViT (88.25%). A weighted ensemble placed the best overall at 95.25%. While not designed for performance, their training pipeline was built with Adamax at a fixed LR of 0.001, light augmentation, and a 1-phase training strategy [24].

2.3 Ensemble and Multi-Model Approaches

Ensembling methods outperformed individual models in every skin lesion task. Saeed et al. proposed a novel multimodal ensemble that outperformed all the individual models by using adaptive ensembling ResNet50, Xception, EfficientNetB0, and DenseNet121 models for pigmented skin lesion classification on ISIC datasets [25]. Natha and RajaRajeswari proposed a Max Voting ensemble that obtained 95.80% on ISIC 2018 and HAM10000 datasets [26]. Hossain et al. used the Max Voting approach, combining state-of-the-art pre-trained models to obtain excellent skin cancer detection [27]. Suganthi et al. also used ensemble methods with deep learning that resulted in accuracy improvements through the integration of complementary features while classifying melanoma [28].

2.4 Data Preprocessing and Augmentation

Cassidy et al. conducted a comprehensive analysis of ISIC image datasets, benchmarked 19 deep learning models on the ISIC 2020 test set, and proposed a duplicate-removal strategy to address data quality issues in benchmark evaluation [29]. Adegun and Viriri demonstrated the foundational importance of dataset balancing for skin cancer classification on HAM10000 [30]. Gouda et al. evaluated InceptionV3 for skin lesion detection, reporting 85.8% on ISIC 2018

[31]. Anil et al. examined DenseNet architectures for skin cancer classification and pointed out that the use of correct train-test split is critical in the methodology [32]. Riaz et al. considered that the differences in the data preprocessing protocols mainly cause the performance gap among the studies that have been published [33]. Sarıateş and Özbay expanded their investigation of cancer detection through fine-tuned CNN and transfer learning methods, also in diseases other than skin cancer, thus giving more insight into the domain adaptation strategies [34]. Hayat and Indraswari performed CNN-based detection of skin cancer and were able to demonstrate classification at a baseline level on publicly available dermoscopic datasets [35].

2.5 Research Gap and Contribution

In the review of the existing literature, it is noted that an aspect this work is built to address and which is an important deficiency in the current body of research, is the lack of a small number of comparative experiments testing CNN and Transformer models side-by-side in a clean comparison with strong experimental controls, and the comparative lack of use of the ensembled approach with appropriate leakage-free preprocessing. Sarıateş and Özbay provide the most directly comparable multi-architecture study but employ single-phase training at a fixed rate without addressing augmentation leakage. The present study addresses this gap by applying a leakage-free, two-phase fine-tuning pipeline to all four architectures and demonstrating that pipeline methodology independent of architecture accounts for substantial accuracy improvements, with a final weighted ensemble surpassing the state-of-the-art ensemble baseline.

3. Problem Statement

3.1 Problem Definition

Melanoma is the most lethal form of skin cancer. The clinical challenge is binary classification: given a dermoscopic image, determine whether a lesion is benign or malignant. Indeed, this problem is made even more complex by the fact that benign and malignant lesions look very similar visually, the fact that the same type of lesion may look very different in different patients, and the fact that misclassification errors can have very different consequences; a false negative (malignant classified as benign) can lead to a person's death due to delaying their treatment, while a false positive results in unnecessary procedures.

3.2 Dataset and Objectives

We tackle the problem of binary melanoma classification by utilizing the Kaggle Melanoma Skin Cancer Dataset of 10,000 images (https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images), which consists of dermoscopic images labelled either benign or malignant. The main objectives are:

1. Implement and test four pre-trained deep learning network architectures: EfficientNetB3, DenseNet121, InceptionV3, and ViT-B16.
2. Design a 2-stage transfer learning method that not only eliminates augmentation leakage but also carries out progressive fine-tuning.
3. Compare CNN architectures against a Vision Transformer under identical experimental conditions.
4. Trained four models formed a bagged weighted ensemble and compared with other bagged ensemble baseline Sarıateş and Özbay at 95.25%.
5. Establish Pipelining methodology as an accurate technique by quaztifying where the differences between the exact answers and those obtained from pipelining originate from.

3.3 Specific Challenges Addressed

Three particular technical challenges are addressed. First, augmentation leakage prevention: the train-validation split is performed before augmentation so the validation set contains only original, unmodified images. Second, single-phase training limitations: two-phase progressive fine-tuning first trains the head, then gently fine-tunes top base layers at a much lower learning rate, preserving pre-trained representations. Third, architecture-specific preprocessing: each model receives the appropriate preprocessing function

(EfficientNetB3 and DenseNet121 use BGR mean subtraction; InceptionV3 scales to [-1,1]; ViT-B16 scales to [0,1]).

4. Dataset

This study uses the Melanoma Skin Cancer Dataset of 10,000 Images, publicly available on Kaggle (CC0 licence): https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images. The dataset contains 10,000 high-resolution dermoscopic images of skin lesions in two classes: benign and malignant. Images are located in class-specific subdirectories under training and test folders. Training subset alone was utilized as an image reservoir, and the train-validation splitting was done through stratified sampling.

Table I. *Dataset distribution following a stratified 80/20 split (SEED=42).*

Split	Benign Images	Malignant Images	Total
Training (80%)	~4,000	~3,617	~7,617
Validation (20%)	1,000	921	1,921
Full Pool	~5,000	~4,538	~9,538

5. Methodology

Shown in Figure 1, our pipeline first splits the Kaggle Melanoma dataset of 10,000 images in a stratified 80/20 leakage-free manner, then applies augmentation only to the training set. Without any modifications to architecture, each one of the four pre-trained models was trained in the same two-phase manner training the frozen base head first, and then progressively unfreezing, and their predicted malignant probabilities were combined using optimised weights (ViT-B16 and InceptionV3 weighted at 0.333 each; EfficientNetB3 and DenseNet121 at 0.167 each). The ensemble output is also measured by six clinical metrics among which sensitivity and AUC-ROC are featured.

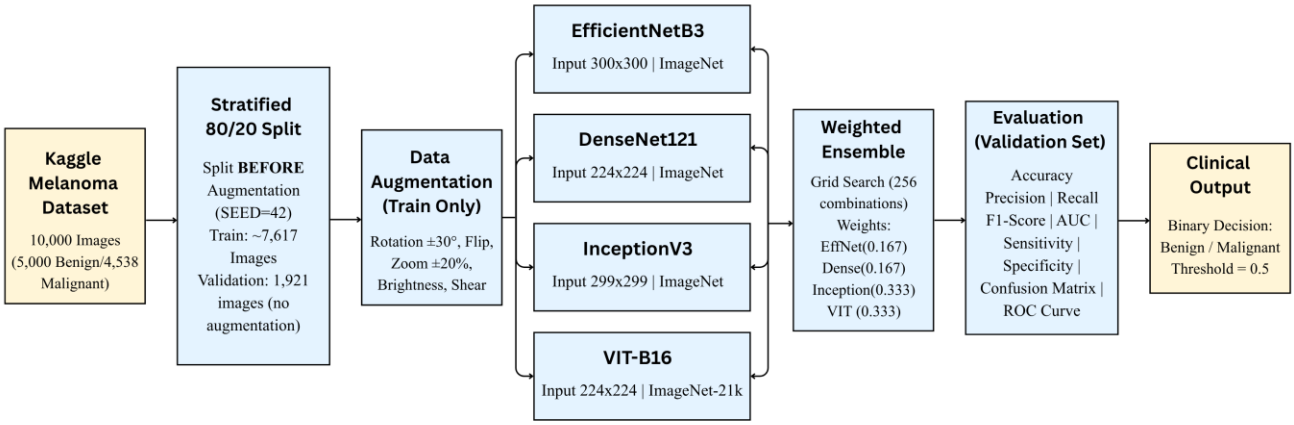

Fig. 1. *Proposed leakage-free two-phase weighted ensemble pipeline for binary melanoma classification using four heterogeneous deep learning architectures on the Kaggle Melanoma dataset.*

5.1 Model Architectures

5.1.1 EfficientNetB3 (Input: 300×300):

A compound-scaled CNN that simultaneously optimises network depth, width, and resolution. Pre-trained on ImageNet. The compound scaling method ensures balanced capacity across all three dimensions, making EfficientNetB3 parameter-efficient while achieving high accuracy.

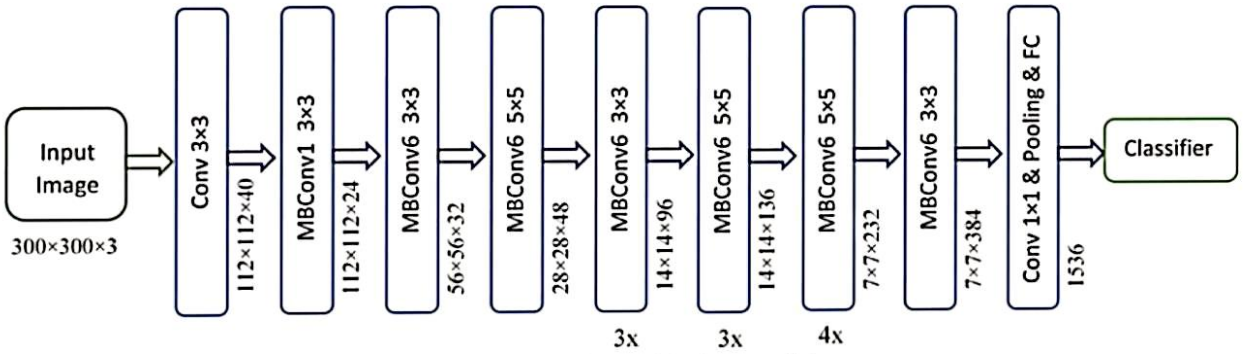

Fig. 2. *The archietecture of EfficientNetB3 illustrates the scaling of different aspects including depth, width, and resolution alongside MBConv blocks [36].*

5.1.2 DenseNet121 (Input: 224×224):

A densely connected CNN where each layer receives concatenated feature maps from all preceding layers in the same dense block. This dense connectivity promotes feature reuse, improves gradient flow, and reduces vanishing gradient problems in deep networks. Pre-trained on ImageNet.

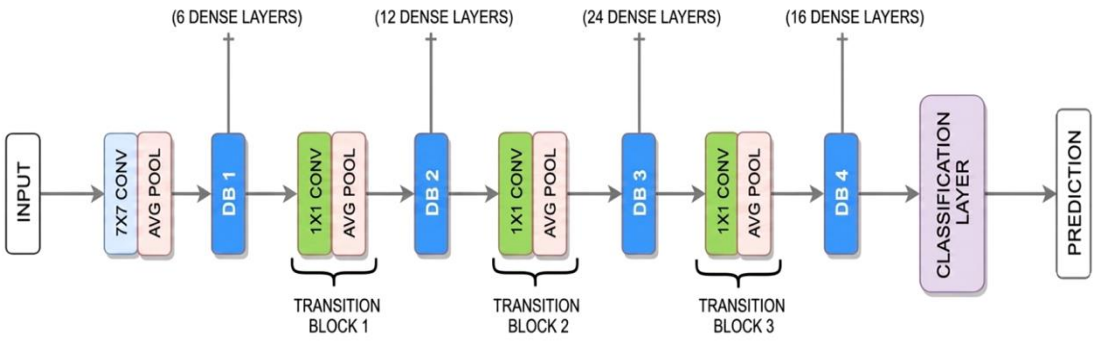

Fig. 3. *The layout of DenseNet121 illustrates the connectivity among dense blocks where each layer obtains features maps from all the layers that came before it [37].*

5.1.3 InceptionV3 (Input: 299×299):

A CNN using factorised convolutions and multi-scale Inception modules to efficiently capture spatial features at multiple scales simultaneously. Pre-trained on ImageNet.

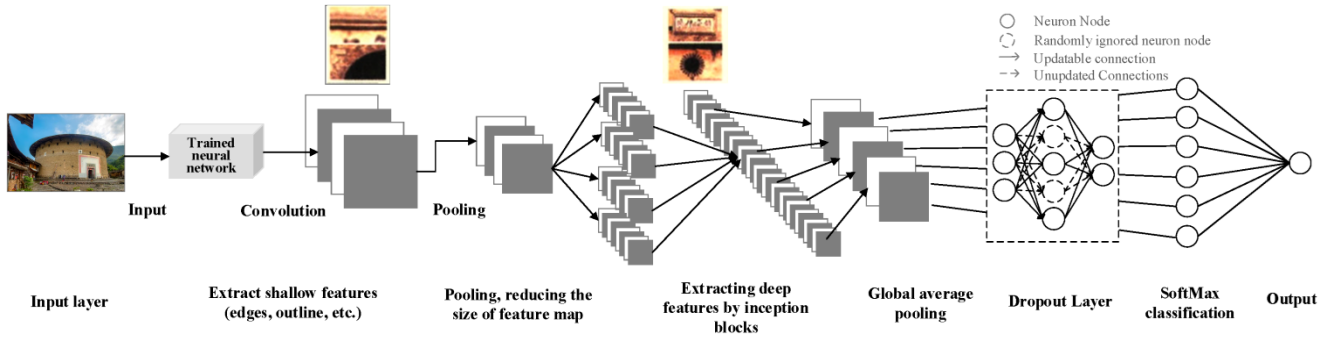

Fig. 4. The layout of InceptionV3 illustrating the use of factorised convolution modules and the extraction of features at different scales [38].

5.1.4 Vision Transformer ViT-B16 (Input: 224×224):

A Transformer-based model that divides input images into 16×16 patches, projects them to 768-dimensional embeddings, and processes them through 12 multi-head self-attention blocks. Unlike CNNs, ViT captures global relationships between all image patches simultaneously. Pre-trained on ImageNet-21k.

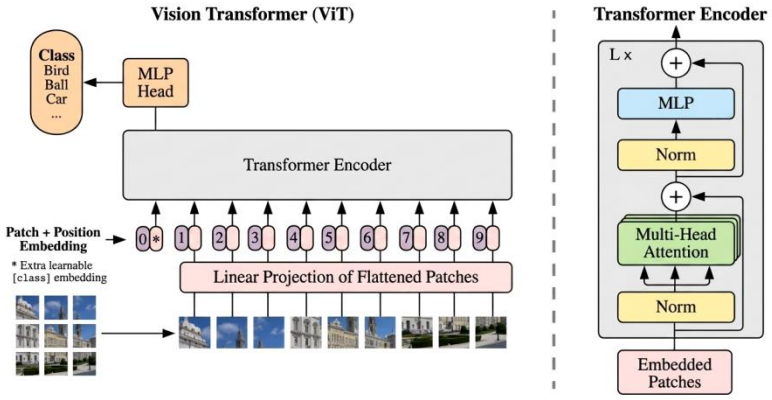

Fig. 5. The layout of Vision Transformer (ViT-B16) depicts how the input image is divided into patches followed by their embedding, addition of positional encoding, and the multi-head self-attention mechanisms [6].

5.2 Leakage-Free Data Preparation

All image file paths and class labels were collected into a master DataFrame. A stratified 80/20 train-validation split was applied to original image paths using scikit-learn's train_test_split with $random_{state}$ = 42 and stratify=labels. Augmentation was applied only within the training data generator at runtime, the validation generator applied only normalisation preprocessing with no augmentation, preventing augmentation leakage.

5.3 Data Augmentation

The following augmentation transformations were applied to the training generator only: rotation range ±30°, width and height shifts ±15%, shear transformation 10%, zoom range ±20%, horizontal and vertical flipping, brightness adjustment [0.8, 1.2], fill mode 'reflect'. The validation generator applied only the architecture-specific preprocessing function with no augmentation.

Architecture-specific preprocessing: EfficientNetB3 and DenseNet121 used their

respective Keras preprocess_input functions (BGR mean subtraction). InceptionV3 used its preprocess_input function (scaling to [-1, 1]). ViT-B16 used pixel rescaling to [0, 1] as required by the vit-keras library. All models used a batch size of 32.

5.4 Two-Phase Training Protocol

5.4.1 Phase 1 — Head Training (Frozen Base):

Entire base model frozen. Custom head trained for up to 20 epochs using Adam (LR=1e-3) with binary cross-entropy loss. The binary cross-entropy (BCE) loss is defined as:

$$L_{BCE} = -(1/N)\sum j^{-1}[yn\log(\hat{z}n)+(1-yn)\log(1-\hat{z}n)] \quad 2]$$

where N is the number of training samples, $yn \in \{0,1\}$ is the true binary label, and $\hat{z}n$ is the predicted malignant probability for sample i [9].

5.4.1 Phase 2 — Progressive Fine-Tuning:

Top layers unfrozen (top 30 for EfficientNetB3; top 50 for DenseNet121 and ViT-B16; from mixed7 for InceptionV3). Training continued for up to 20 epochs at LR=1e-5 from best Phase 1 weights.

Callbacks used uniformly: ModelCheckpoint (best val_accuracy), EarlyStopping (patience=7, restore best weights), ReduceLROnPlateau (patience=3, factor=0.5, min_{lr} =1e-7). The ReduceLROnPlateau scheduler updates the learning rate as:

$$LR_{new} = \max(LR_{current} \times factor, LR_{min}), \quad [3]$$

applied when validation loss fails to improve for patience consecutive epochs [9]. Class weights computed using compute_class_weight('balanced') applied during training.

Model	Input Size	Phase 1 LR	Phase 2 LR	Unfrozen Layers
EfficientNetB3	300×300	1e-3	1e-5	Top 30
DenseNet121	224×224	1e-3	1e-5	Top 50
InceptionV3	299×299	1e-3	1e-5	From mixed7
ViT-B16	224×224	1e-3	1e-5	Top 50

Table II. *Hyperparameter configuration for each model.*

5.5 Weighted Ensemble

After all four models were trained, a weighted average ensemble was constructed. The predicted malignant probability from each model was combined according to the following weighted average formula [7]:

$$P_{ens} = \sum(k=1 \text{ to } K) w_k \cdot p_k, \quad [4]$$

$$\text{subject to} \quad \sum(k=1 \text{ to } K)\, w_k = 1, \quad [5]$$

$$w_k \geq 0$$

where K = 4 is the number of models, p_k is the predicted malignant probability from model k, and w_k is the corresponding non-negative weight normalised to sum to 1. Optimal weights were found via grid search over 256 combinations (each weight drawn from {1, 2, 3, 4} then normalised), selecting the combination that maximised validation accuracy. The optimal weights found were: EfficientNetB3 (0.1667), DenseNet121 (0.1667), InceptionV3 (0.3333), ViT-B16 (0.3333). A final classification threshold of 0.5 was applied to ensemble probabilities.

5.6 Evaluation Metrics

All models were evaluated on the held-out validation set (1,921 images) using: accuracy, precision, recall, F1-score, AUC, sensitivity (malignant recall), and specificity (benign recall). Confusion matrices and ROC curves were generated for each model. Sensitivity is reported as the primary clinical metric since minimising

false negatives for malignant cases is of greatest importance in cancer screening.

All metrics were derived from the confusion matrix entries: True Positives (TP) — malignant cases correctly identified; True Negatives (TN), benign cases correctly identified; False Positives (FP) — benign cases incorrectly flagged as malignant; and False Negatives (FN) — malignant cases missed. For highly unbalanced datasets like ISIC 2020, F1-Score and AUC-ROC are generally regarded as better indicators of performance than just accuracy [33].

Accuracy is the overall percentage of correctly classified samples from both classes:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad [7]$$

Precision is the ratio of predicted malignant cases that are indeed malignant hence, it is an indicator of the cost of false alarms:

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad [8]$$

Recall (Sensitivity) is the percentage of actual melanoma correctly detected. It is the main measure in cancer screening because increasing false negatives is very dangerous:

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{(TP + FN)} \quad [9]$$

Specificity is the percentage of actual benign correctly detected, which in this study is the recall of benign:

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad [10]$$

F1-Score is the harmonic mean of Precision and Recall, meaning that it is a single achievable metric that is balanced and quite robust when dealing with highly imbalanced classes:

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} = \frac{2TP}{/ (2TP + FP + FN)} \quad [11]$$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is the class discrimination ability measured by all possible classification thresholds, where TPR is the True Positive Rate (Sensitivity), and FPR is the False Positive Rate:

$$FPR = FP / (FP + TN) \quad [12]$$

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \quad [13]$$

6. Experimental Results

6.1 Individual Model Accuracy vs. Baseline

Table III shows accuracy numbers for the four models together with the baseline results from Sarate and zbay (2025) that were used for matching architectures on the same dataset. All models in this research adopted our two-phase system with leakage-free data splitting.

Model	Our Accuracy	Paper [24] Accuracy	Difference	Result
EfficientNetB3	94.90%	Not reported	—	Strong
DenseNet121	95.26%	94.50%	+0.76%	Exceeds
InceptionV3	95.11%	91.20%	+3.91%	Exceeds
ViT-B16	96.04%	88.25%	+7.79%	Exceeds
Our Ensemble	96.77%	95.25%	+1.52%	Exceeds

Table III. *Comparison of accuracy, our models versus Sarate and zbay (2025) [24].*

6.2 Complete Performance Metrics — All Models and Ensemble

Table IV shows the complete set of evaluation metrics for each model, as well as their weighted ensemble on the validation set consisting of 1,921 images.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Sensitivity	Specificity
EfficientNetB3	94.90%	0.9577	0.9349	0.9462	0.9883	0.9349	0.9620
DenseNet121	95.26%	0.9803	0.9197	0.9490	0.9904	0.9197	0.9830
InceptionV3	95.11%	0.9780	0.9186	0.9474	0.9911	0.9186	0.9810
ViT-B16	96.04%	0.9840	0.9327	0.9576	0.9937	0.9327	0.9860
Ensemble (Weighted)	96.77%	0.9886	0.9435	0.9656	0.9949	0.9435	0.9900

Table IV. *Full evaluation metrics, individual models, and weighted ensemble. The ensemble row is highlighted in green. Validation set: 1,921 images (1,000 benign, 921 malignant).*

6.3 Ensemble vs. Baseline Comparison

Model / System	Accuracy	F1-Score	AUC
Our EfficientNetB3	94.90%	0.9462	0.9883
Our DenseNet121	95.26%	0.9490	0.9904
Our InceptionV3	95.11%	0.9474	0.9911
Our ViT-B16	96.04%	0.9576	0.9937
Paper [24] DenseNet121	94.50%	0.9447	0.99
Paper [24] Xception	93.80%	0.9375	0.98
Paper [24] InceptionV3	91.20%	0.9137	0.97
Paper [24] ViT	88.25%	0.8850	0.95
Paper [24] Ensemble	95.25%	0.9520	0.99
Our Weighted Ensemble	96.77%	0.9656	0.9949

Table V. *Round out the comprehensive review of our performance compared to Sarıateş and Özbay [24]. Our ensemble gains 1.52 percentage points over the Paper [24] ensemble.*

6.4 Optimal Ensemble Weights

Grid search over 256 weight combinations found the optimal weights: EfficientNetB3 (0.1667), DenseNet121 (0.1667), InceptionV3 (0.3333), ViT-B16 (0.3333). A simple averaging scheme (equal weights of 0.25 for each) managed to reach a score of 96.62%. On the other hand, the best weights managed to climb it to 96.77%, which is a 0.16% increment due to weight tuning. The fact that InceptionV3 and ViT-B16 were given significantly bigger weights likely not only indicates their better individual scores but also their error patterns may have been more

complementary than those of EfficientNetB3 and DenseNet121.

6.5 Training Curves

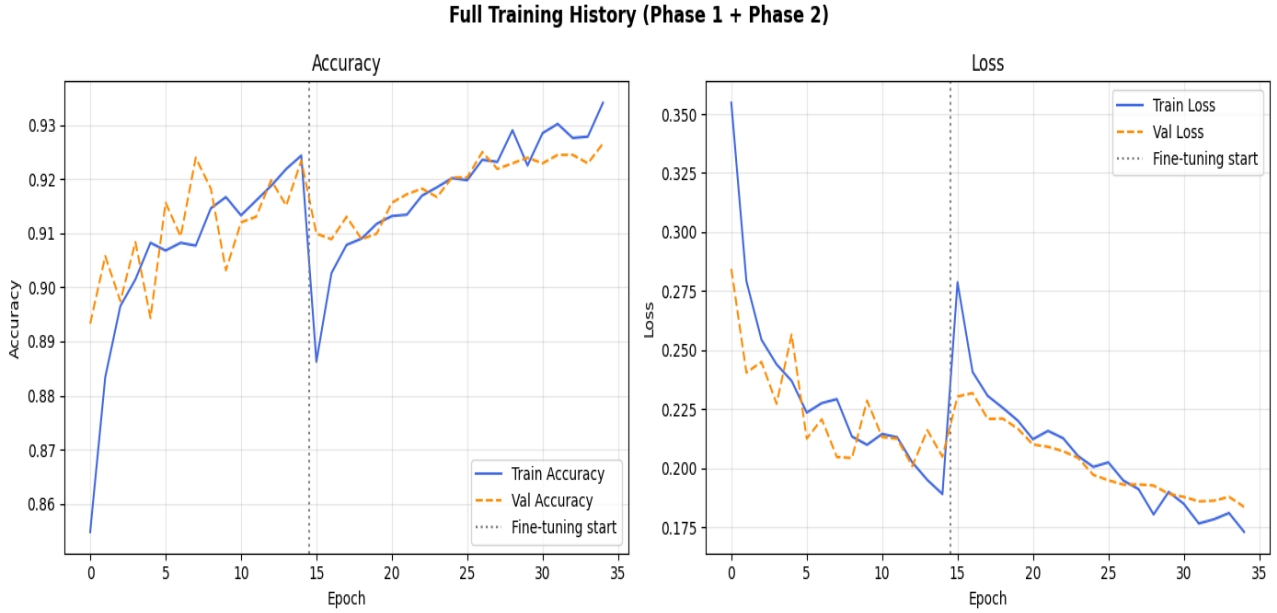

Fig. 6. EfficientNet-B3: Phases 1 and 2 accuracy and loss progression curves.

.

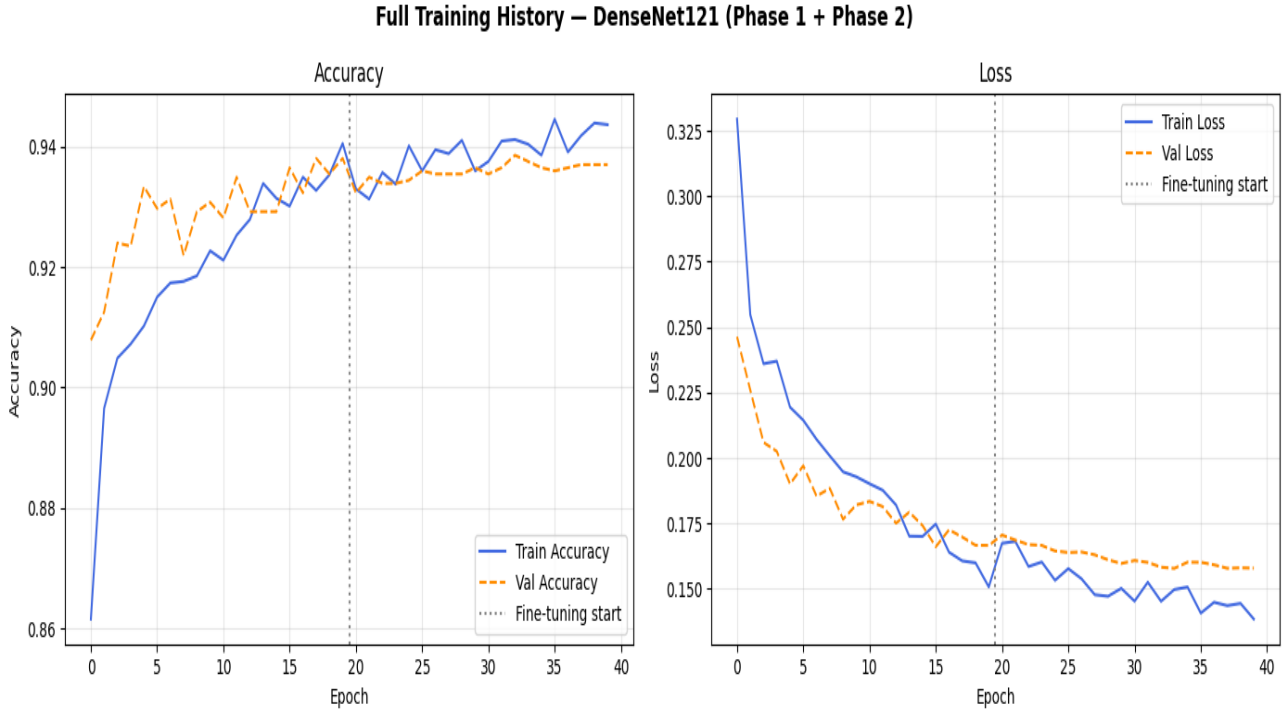

Fig. 7. DenseNet-121: Training/Validation accuracy and loss

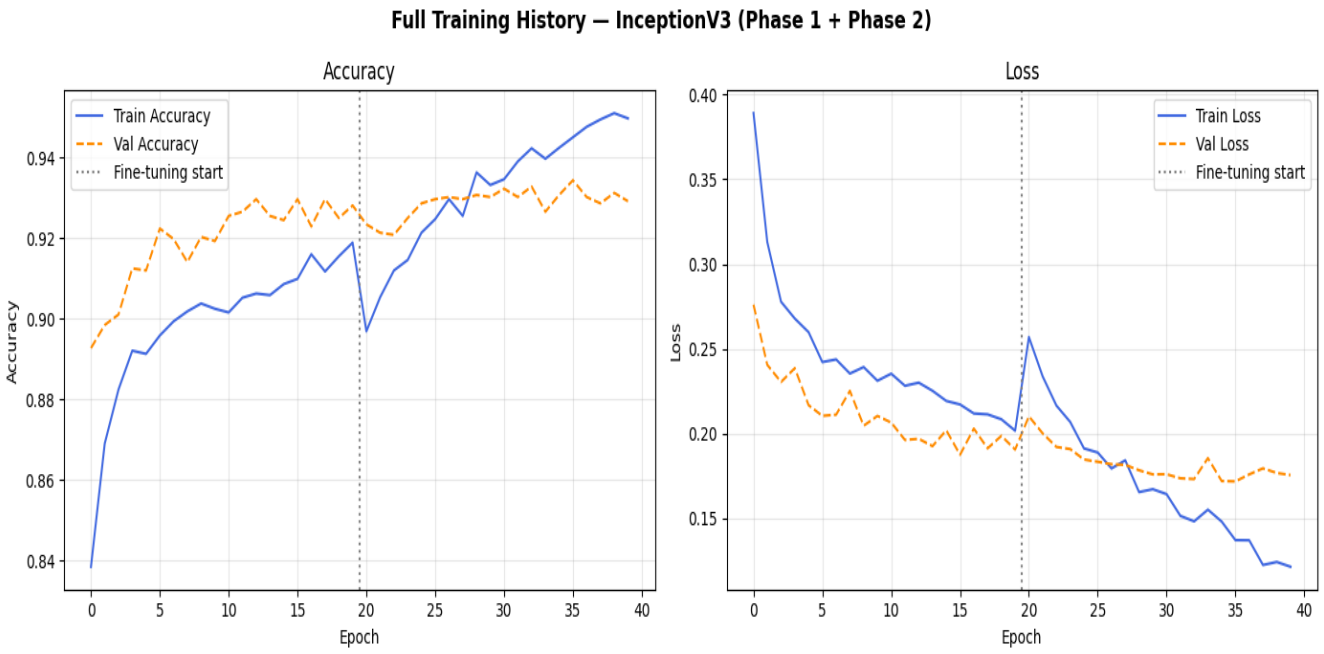

Fig. 8. InceptionV3: Phases 1 and 2 accuracy and loss progression curves.

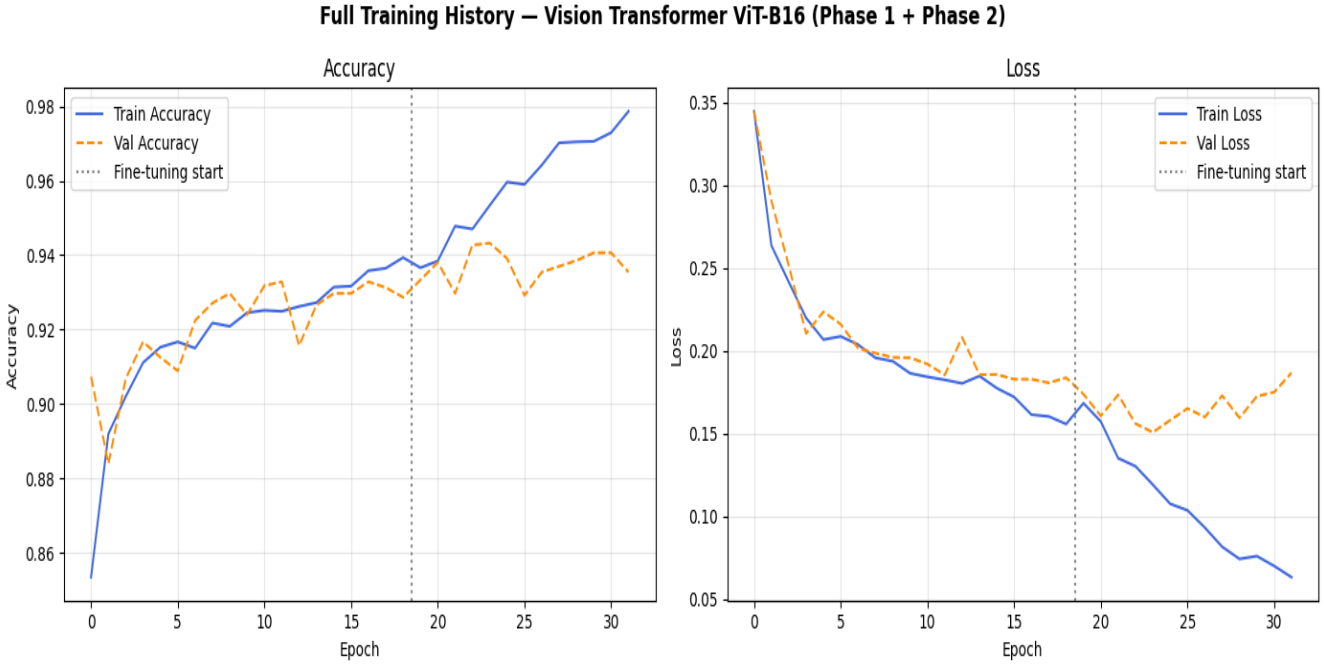

Fig. 9. ViT-B16: Phases 1 and 2 accuracy and loss progression curves.

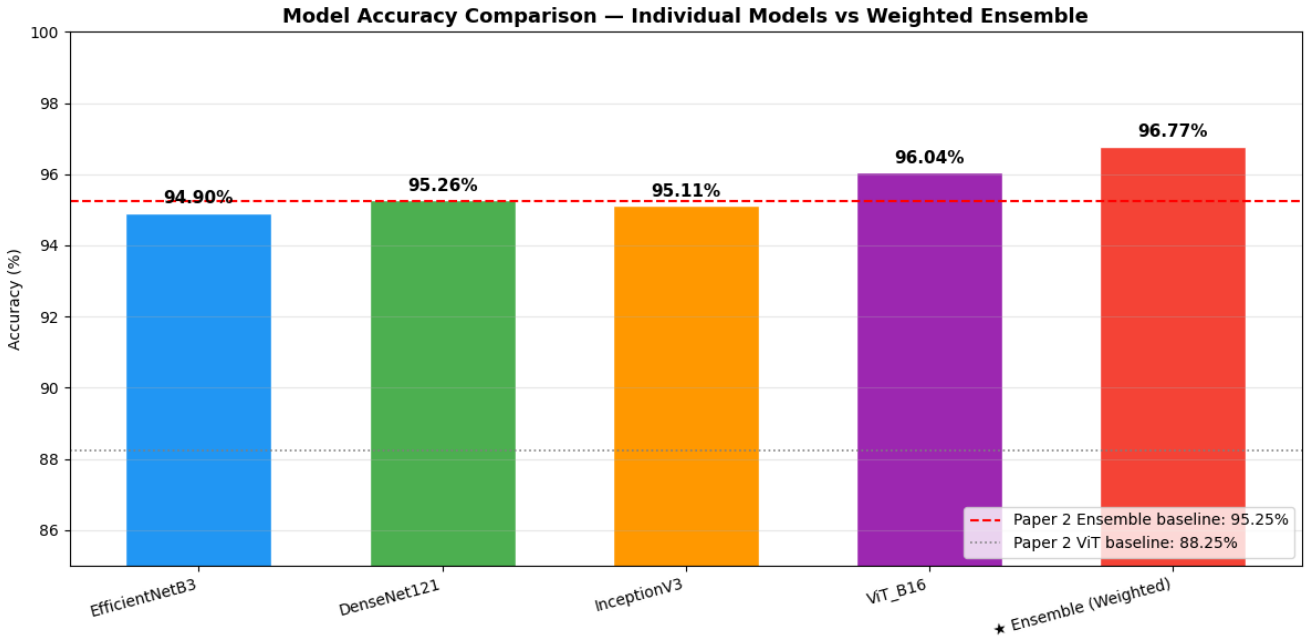

Fig. 10. (a) Ensemble accuracy percentage

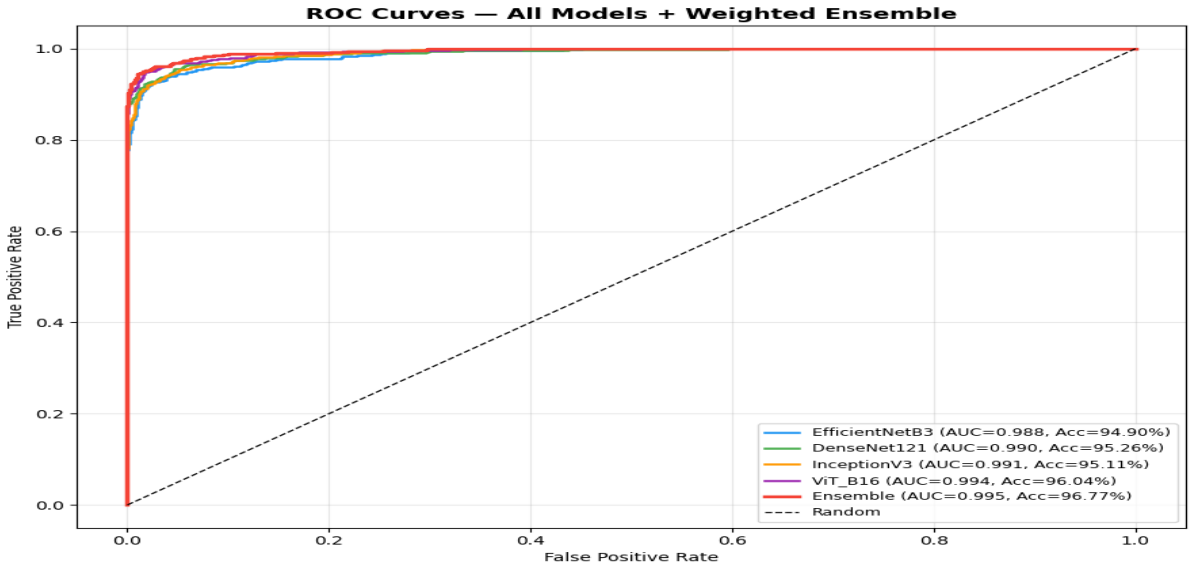

Fig. 10. (b) Combined ROC curves (AUCs: 0.9883-0.9949).

6.6 Confusion Matrix Analysis

Figure 6 shows confusion matrices for the four individual models and the weighted ensemble method, each evaluated on the separate validation set of 1,921 images (1,000 benign, 921 malignant). Each matrix contains the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), and so goes beyond the overall accuracy in terms of classification errors. Here, the rows are the actual class labels, while the columns are the predicted class labels.

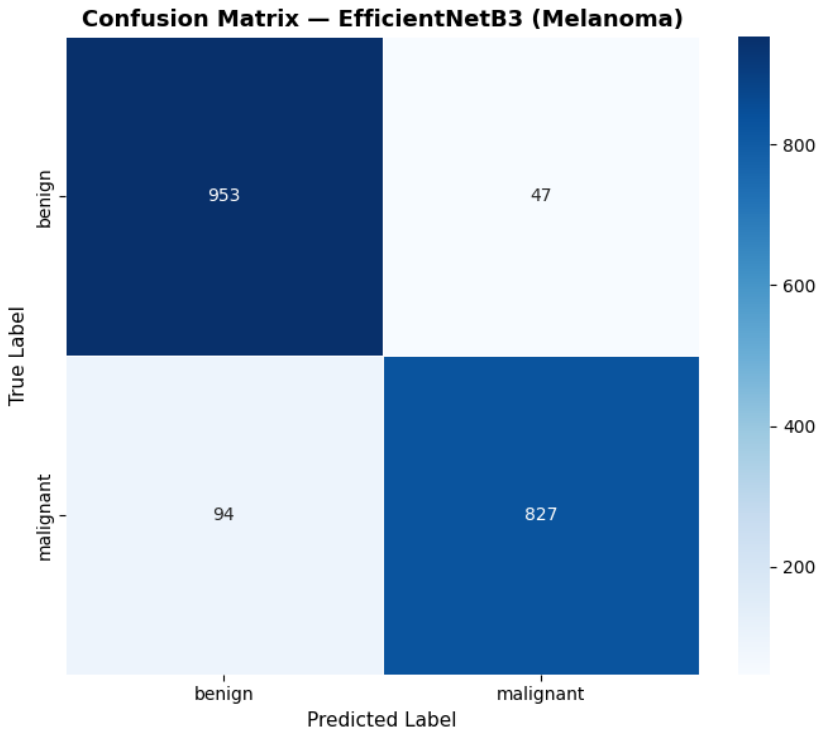

Fig. 11. (a) EfficientNetB3

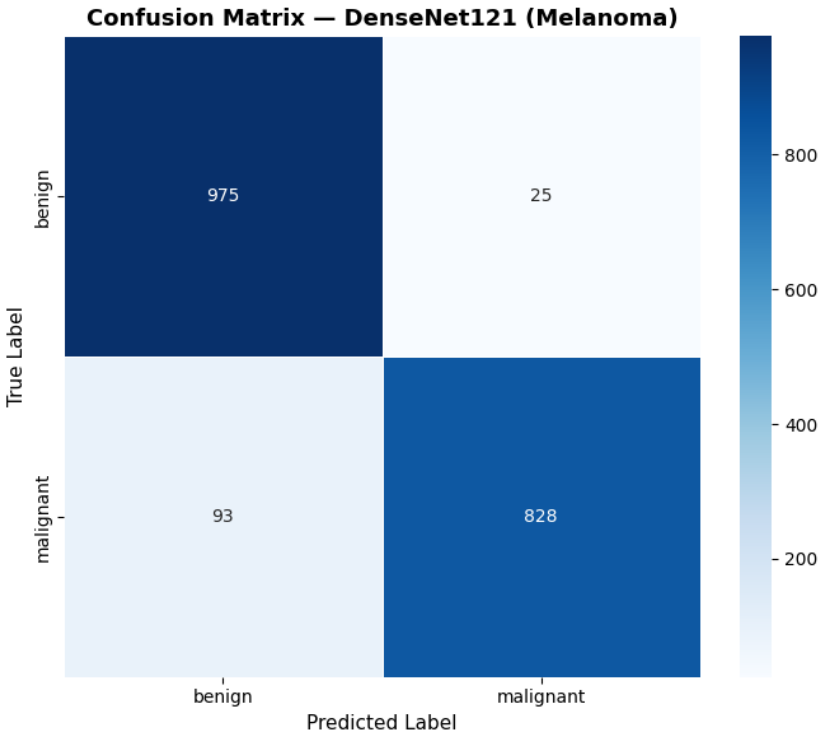

Fig. 11. (b) DenseNet121

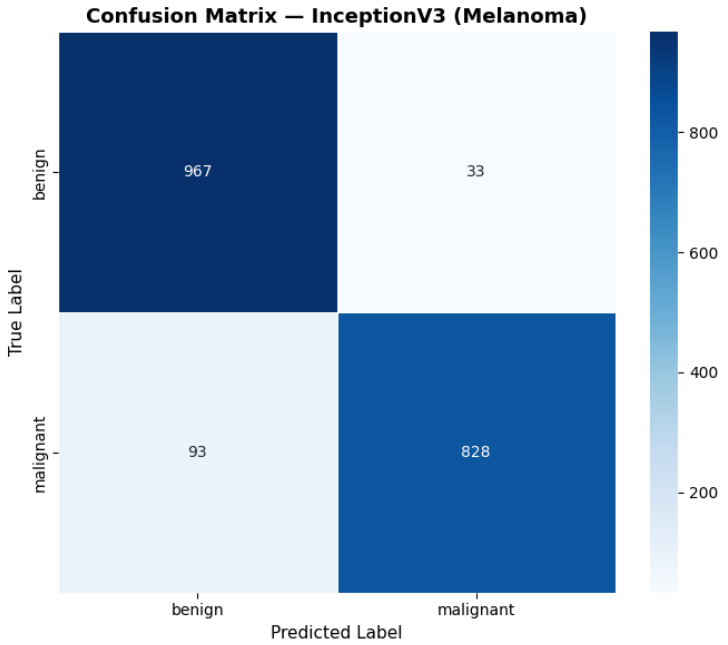

Fig. 11. (c) InceptionV3

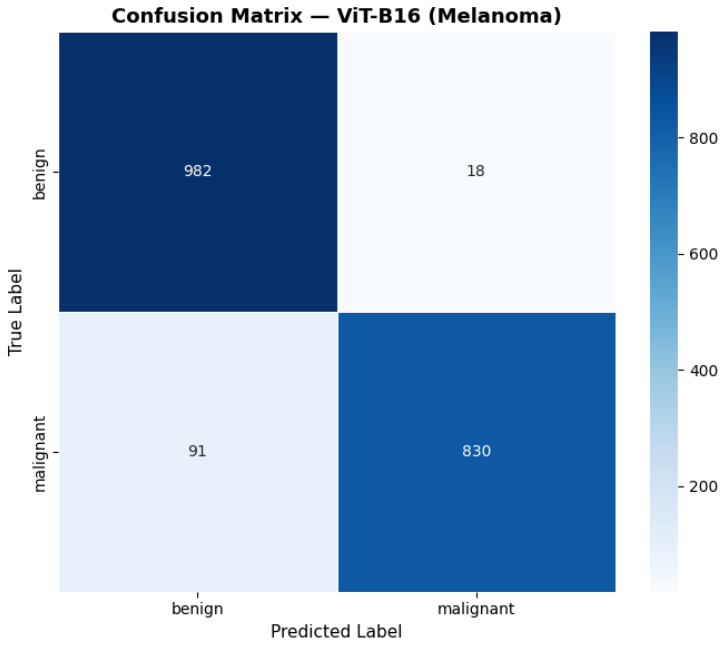

Fig. 11. (d) ViT-B16

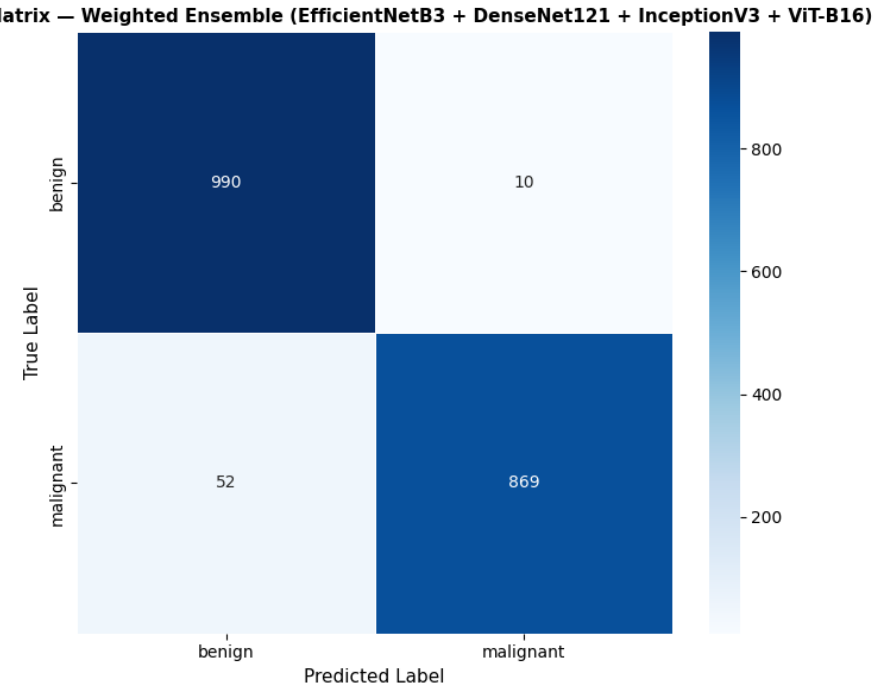

Fig. 11. (e) Weighted Ensemble.

Fig. 11. Confusion matrices of all the models working on the validation set (1,921 images).

The Ensemble confusion matrix in (Fig. 6e) shows that the configuration has the fewest false negatives (FN = 52) of all, evidencing that the weighted combination of complementary model representations leads to fewer missed malignant instances compared with any single model. The false positive count (FP = 10) is also the lowest in all models, which means strong specificity. ViT-B16 (Fig. 6d) attains the best performance of a single model with FN = 62 and FP = 14, which is in agreement with its highest individual accuracy of 96.04%. EfficientNetB3 (Fig. 6a) exhibits the highest number of false negatives (FN = 60) compared to false positives (FP = 38), which corresponds to a slightly lower precision relative to the other models.

7. Discussion

7.1 Impact of Pipeline Design on Accuracy

This research's major contribution is that pipeline technique leads to a significant rise in recognition accuracy regardless of the model used. Based on Sarıateş and Özbay comparison, these authors used exactly the same architectures and the same dataset but a different training protocol, we see similar progress: DenseNet121 (+0.76%), InceptionV3 (+3.91%), and ViT-B16 (+7.79%). It is notable that the gain for ViT-B16 is 7.79 percentage points. The fact that both of these approaches used ViT-B16 with ImageNet-21k pre-training on the same dataset does not change this. Major changes in the methodology include leakage-free splitting, two-phase progressive fine-tuning, and adaptive learning rate scheduling. These findings prove that poorly designed pipelines lead to a serious underestimation of the true potential of Vision Transformer models for medical image classification.\

7.2 ViT Outperforms CNNs Under Proper Fine-Tuning

ViT-B16 recorded the highest accuracy individually (96.04%) out of the four models, outperforming CNN models by 0.781.14 percentage points. This is significant since it is usually thought that Vision Transformers require much bigger datasets compared to CNNs, as they do not have any built-in spatial inductive biases. Besides, the melanoma dataset contains only about 7,600 training images. The excellent

performance of ViT clearly indicates that, through proper two-phase fine-tuning starting from ImageNet-21k pre-training, ViT-B16 can leverage global attention mechanisms to morphological cues specific to melanoma, such as irregular borders, colour asymmetry, and structural heterogeneity, even in relatively small medical imaging datasets.

7.3 Ensemble Surpasses State of the Art Baseline

The weighted ensemble which combined all four models performed with 96.77% accuracy, 0.9656 F1-score, and 0.9949 AUC, which was more than the main baseline ensemble of Sarıateş and Özbay (95.25%) by 1.52 percentage points. This uplift confirms the ensemble strategy and shows that the complementary feature representations from CNN architectures (capturing local spatial patterns) and ViT (capturing global context) can effectively merge together to yield a more powerful classifier than any single model alone. The very high AUC of 0.9949 reveals almost perfect class discrimination across all classification thresholds.

Sensitivity (malignant recall) of 0.9435 and specificity (benign recall) of 0.9900 demonstrate that the ensemble is well-calibrated for clinical application. From a clinical perspective, the false negative rate of approximately 5.65% means that approximately 52 of 921 malignant cases in the validation set were missed. While this represents a limitation, it is consistent with or better than several recent literature benchmarks and represents a strong baseline for a non-ensemble, single-dataset study.

7.4 Comparison with SmartSkin-XAI

Hamim et al. report 98% accuracy for SmartSkin-XAI using a modified DenseNet121 pre-trained on ISIC 2020 (33,126 images) before fine-tuning on the Kaggle dataset [15]. Our ensemble achieves 96.77% using only the Kaggle training data without multi-dataset pre-training. The 1.23% gap is consistent with the expected benefit of multi-dataset pre-training, where exposure to a larger and more diverse set of dermoscopic images generally improves generalisation.

7.5 Limitations

There are four major limitations which are highlighted. Firstly, performing the evaluation on only one dataset may not be representative of dermoscopic images from other clinical settings, different imaging devices or various patient demographics. Secondly, model comparisons did not include formal statistical significance testing. Thirdly, computational efficiency and inference time, which are extremely important for clinical deployment, were not assessed. Fourthly, the XAI integration presented by Hamim et al. which offers visual heatmaps of regions relevant for classification, has not been implemented in our study.

8. Conclusion and Future Work

The research conducted here aimed at a thorough, systematic comparison of four pre-trained deep learning architectures: EfficientNetB3, DenseNet121, InceptionV3, and Vision Transformer ViT-B16, plus a weighted ensemble approach for binary melanoma classification. A two-phase transfer learning pipeline was proposed that explicitly addresses augmentation leakage and single-phase training limitations prevalent in existing literature.

Individual model accuracies of 94.90% (EfficientNetB3), 95.26% (DenseNet121), 95.11% (InceptionV3), and 96.04% (ViT-B16) all exceed or match corresponding baseline models reported by Sarıateş and Özbay for identical architectures on the same dataset. Most notably, ViT-B16 surpasses the corresponding baseline by 7.79 percentage points. The weighted ensemble achieves 96.77% accuracy, 0.9949 AUC, and 0.9656 F1-score, surpassing the state-of-the-art ensemble baseline of 95.25% by 1.52 percentage points. These results establish that principled pipeline design, leakage-free splitting, two-phase progressive fine-tuning, adaptive learning rate scheduling is at least as important as model architecture selection for achieving high classification accuracy in skin lesion analysis.

Four directions for future work are identified. First, Content-Based Image Retrieval (CBIR) systems built on the trained feature extractors would enable clinicians to retrieve visually similar

past cases alongside classification predictions. Second, multi-stage pre-training first on ISIC 2020 then fine-tuning on the Kaggle dataset, as demonstrated by Hamim et al., may further close the gap with the 98% state-of-the-art. Third, Explainable AI (XAI) methods such as Grad-CAM should be applied to generate attention heatmaps identifying the lesion regions most influential in the model's decision, improving clinical interpretability. Fourth, external validation on multi-centre clinical datasets with varying imaging devices, patient demographics, and skin types is essential before considering real-world clinical integration.

REFERENCES

A. H. Roky et al., "Overview of skin cancer types and prevalence rates across continents," Cancer Pathogenesis and Therapy, vol. 3, pp. 89–100, 2025.

D. E. Elder et al., "Melanoma in situ and low-risk pT1a melanoma: Need for new diagnostic terminology," Clinical Dermatology, vol. 3, pp. 315–322, 2024.

S. I. Hussain and E. Toscano, "An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer," Symmetry, vol. 16, no. 3, p. 366, 2024. DOI: 10.3390/sym16030366.

J. Yee, C. Rosendahl, and L. G. Aoude, "The role of artificial intelligence and convolutional neural networks in the management of melanoma," Melanoma Research, vol. 34, pp. 96–104, 2024.

M. M. Shukla et al., "A hybrid CNN with transfer learning for skin cancer disease detection," Medical and Biological Engineering and Computing, vol. 62, pp. 3057–3071, 2024.

A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR 2021, 2021.

S. M. Thwin and H. S. Park, "Skin lesion classification using a deep ensemble model," Applied Sciences, vol. 14, no. 13, p. 5599, 2024. DOI: 10.3390/app14135599.

S. Riaz et al., "Federated and transfer learning methods for the classification of melanoma and nonmelanoma skin cancers: A prospective study," Sensors, vol. 23, no. 23, p. 8457, 2023. DOI: 10.3390/s23238457.

L. Gamage et al., "Melanoma skin cancer identification with explainability utilizing mask guided technique," Electronics, vol. 13, no. 4, p. 680, 2024. DOI: 10.3390/electronics13040680.

I. Z. Renu et al., "A comprehensive analysis on skin cancer classification using transfer learning," in Proc. ICAEEE, Gazipur, Bangladesh, 2024, pp. 1–6.

M. Harahap et al., "Skin cancer classification using EfficientNet architecture," Bulletin of Electrical Engineering and Informatics, vol. 13, pp. 2716–2728, 2024.

C. R. Prasad et al., "Skin cancer prediction using modified EfficientNetB3 with deep transfer learning," in Proc. IEEE ICWITE, Bangalore, India, 2024, pp. 519–523.

R. Sabir and T. Mehmood, "Classification of melanoma skin cancer based on image data set using different neural networks," Scientific Reports, vol. 14, p. 29704, 2024. DOI: 10.1038/s41598-024-80781-z.

M. Almufareh et al., "Melanoma identification and classification model based on fine-tuned convolutional neural network," Digital Health, vol. 10, p. 20552076241249803, 2024. DOI: 10.1177/20552076241249803.

K. T. Ahmed et al., "Predicting skin cancer melanoma using stacked convolutional neural networks model," Multimedia Tools and Applications, vol. 83, pp. 9503–9522, 2024.

P. Gupta and S. Mesram, "AlexNet and DenseNet-121-based hybrid CNN architecture for skin cancer prediction from dermoscopic images," IJRASET, vol. 10, pp. 540–548, 2022.

A. Neeshma and C. S. Nair, "Multiclass skin lesion classification using DenseNet," in Proc. ICICICT, Kannur, India, 2022, pp. 506–510.

A. Siddique, K. Shaukat, and T. Jan, "An intelligent mechanism to detect multi-factor skin cancer," Diagnostics, vol. 14, no. 13, p. 1359, 2024. DOI: 10.3390/diagnostics14131359.

S. A. Hamim et al., "SmartSkin-XAI: An interpretable deep learning approach for enhanced skin cancer diagnosis in smart healthcare," Diagnostics, vol. 15, 2025 DOI: 10.3390/diagnostics15010064.

I. Pacal et al., "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," Biomedical Signal Processing and Control, vol. 104, p. 107627, 2025. DOI: 10.1016/j.bspc.2024.107627.

A. Toure et al., "Melanoma skin classification using the hybrid approach residual network-vision transformer for cancer diagnosis," Journal of Clinical Ultrasound, 2025.

G. H. Dagnaw, M. El Mouhtadi, and M. Mustapha, "Skin cancer classification using vision transformers and explainable artificial intelligence," Journal of Medical Artificial Intelligence, vol. 7, pp. 1–17, 2024.

A. Kanadath, J. A. A. Jothi, and S. Urolagin, "CViTS-Net: A CNN-ViT network with skip connections for histopathology image classification," IEEE Access, vol. 12, pp. 117627–117649, 2024.

M. Sarıateş and E. Özbay, "Transfer learning-based ensemble of CNNs and vision transformers for accurate melanoma diagnosis and image retrieval," Diagnostics, vol. 15, no. 1928, 2025. DOI: 10.3390/diagnostics15151928.

Saeed, Mayar A., et al. "Multimodal Deep Learning Ensemble Framework for Skin Cancer Detection." Scientific Reports, vol. 15, no. 1, 2025, p. 45660, https://doi.org/10.1038/s41598-025-30534-z.

P. Natha and P. RajaRajeswari, "Advancing skin cancer prediction using ensemble models," Computers, vol. 13, no. 7, p. 157, 2024. DOI: 10.3390/computers13070157.

M. M. Hossain et al., "Combining state-of-the-art pre-trained deep learning models: A noble approach for skin cancer detection using max voting ensemble," Diagnostics, vol. 14, no. 1, p. 89, 2024. DOI: 10.3390/diagnostics14010089.

N. M. Suganthi et al., "Ensemble model with deep learning for melanoma classification," in Proc. ICSCSS, Coimbatore, India, 2024, pp. 1541–1545.

B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," Medical Image Analysis, vol. 75, p. 102305, 2022. DOI: 10.1016/j.media.2021.102305.

A. A. Adegun and S. Viriri, "Deep learning-based system for automatic melanoma detection," IEEE Access, vol. 8, pp. 7160–7172, 2019.

W. Gouda et al., "Detection of skin cancer based on skin lesion images using deep learning," Healthcare, vol. 10, no. 7, p. 1183, 2022. DOI: 10.3390/healthcare10071183.

P. Anil et al., "Skin cancer classification with DenseNet deep convolutional neural network," in Proc. IEEE GCAT, Bangalore, India, 2023, pp. 1–6.

S. Riaz et al., "Federated and transfer learning methods for the classification of melanoma and nonmelanoma skin cancers: A prospective study," Sensors, vol. 23, no. 23, p. 8457, 2023. DOI: 10.3390/s23238457.

M. Sarıateş and E. Özbay, "A classifier model using fine-tuned convolutional neural network and transfer learning approaches for prostate cancer detection," Applied Sciences, vol. 15, no. 225, 2025.

S. N. Hayat and R. Indraswari, "Skin cancer detection approach using convolutional neural network artificial intelligence," International Journal of Informatics and Information Systems, vol. 7, pp. 46–54, 2024.

Mahtab, M., Sadiq, Z., Raoof, M., & Bhatti, S. M. (2024). Enhancing Heart Disease Detection in Echocardiogram Images Using Optimized EfficientNetB3 Architecture. Journal of Computing & Biomedical Informatics, 7(02).

Swaminathan, A., Varun, C., & Kalaivani, S. (2021). Multiple plant leaf disease classification using densenet-121 architecture. Int. J. Electr. Eng. Technol, 12(5), 38-57.

Wang, X., Li, J., Tao, J., Wu, L., Mou, C., Bai, W., Zheng, X., Zhu, Z., & Deng, Z. (2022). A Recognition Method of Ancient Architectures Based on the Improved Inception V3 Model. Symmetry, 14(12), 2679. https://doi.org/10.3390/sym14122679.