International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 1-18
https://doi.org/10.70917/ijcisim-2026-2864
Article

Vision Transformer Outperforms CNN Architectures in Binary Skin Lesion Classification: An Eight-Model Controlled Deep Learning Study

Bharti G. Gadge[1], Vikul J. Pawar[2], Vinod Damdhar[3], Kailash Kharat[4]

[1,2]Department of Computer Science and Engineering, Government College of Engineering Aurangabad, Chhatrapati Sambhajinagar, M.S. India-431001

[3,4]CSMSS, Chh. Shahu College of Engineering, Kanchanwadi, Chhatrapati Sambhajinagar, M.S. India-431011

*Corresponding Author: damdharbv@gmail.com

Abstract: —Skin cancer is one of the life-threatening malignancies in the world, timely detection of which is directly proportional to the survival of patients. Traditional dermoscopic diagnosis suffers inter-observer variability, lack of specialists and is not scalable in resource limited healthcare environments. The end-to-end hierarchical feature learning provided by deep learning is a transformative solution to the learned dermoscopic image corpora. The paper empirically comparatively examines eight binary skin lesion classifiers benign versus malignant of a dataset of 2,637 training images, and 661 held out test images. The assessed architectures are located on a wide design range a custom CNN trained using fresh data, two ResNet18 transfer learning pipelines, DenseNet121, MobileNetV3-Large, ViT-Small/16 (ImageNet-21K), ConvNeXt-Tiny (ImageNet-12K) and EfficientNetV2-S (ImageNet-21K). All the models are trained with the same conditions involving stratified splitting, Weighted Random Sampler, two-stage fine tuning with discriminative learning rates, Automatic Mixed Precision, and early stopping. It is evaluated using six metrics accuracy, per-class precision, recall, F1-score, ROC-AUC, and PR-AUC. The highest test accuracy (91.53) and macro-F1 (0.907) is attained with EfficientNetV2-S. ViT-Small/16 has the best ROC-AUC (0.9723) and PR-AUC (0.9699), which proves the effectiveness of Vision Transformer in threshold-free probabilistic discrimination the clinically decisive measure in screening applications. The three contemporary timm-based models have consistently reached ROC-AUC 0.95 and above, but the legacy CNN models are at 0.56 even though the legacy CNN models are at competitive accuracy of 89-90%. MobileNetV3-Large yields a false negative rate of 61 (20% miss rate), which highlights the clinical risk of aggressive model compression. The findings simplify the selection of the model in clinical studies in implementing dermoscopy, suggesting that ViT-Small/16 should be used as a probabilistic-ranked malignancy screening model, and EfficientNetV2-S should be used as a fixed threshold binary triage model in a telemedicine system.

Keywords: —Skin Lesion Classification, Binary Dermoscopic Analysis, Vision Transformer (ViT), EfficientNetV2, Transfer Learning, ROC-AUC, Convolutional Neural Network.

1. INTRODUCTION

One of the most common types of cancer in the world is skin cancer with the World Health Organization estimating that every one out of every five people will be affected by any type of skin cancer at one point in their lives. The most clinically hazardous form, melanoma, contributes most to the number of deaths caused by skin cancer. Its prognosis is heavily stage-dependent. five-year survival is more than 98% that of localized lesions but only about 30% once the disease is metastatic, so early detection is directly life-saving. Traditional diagnosis is based on dermoscopy a non-invasive optical imaging that enhances diagnostic sensitivity by 10-27% of naked-eye examination but demands

intensive training of specialists, inter-observer variation has a lengthy history of high variability, and is not available in low and middle-income countries where specialist shortages are the greatest. The structural constraints present a strong demand to AI-supported devices that can screen on a regular basis with high accuracy regardless of the availability of specialists.

Deep learning has changed the analysis of medical images by end-to-end learning hierarchical feature representations. Esteva et al. (2017)[1] landmark study has shown CNN level of performance as a dermatologist when it comes to 757 skin disease categories, which is a clinically credible diagnostic value. Later innovations ResNet residual connections[2], DenseNet dense inter layer reuse[3], EfficientNet compound scaling[12], and MobileNet depthwise separable convolutions allowed more and more powerful generalization of limited dermoscopic training data using ImageNet transfer learning. Vision transformers (ViTs)[10] and CNN, more recently, replaced CNN hegemony by learning long-range spatial features on a patch basis with self-attention and CNN-based designs, including ConvNeXt[11], crossed the CNN-ViT design gap. Nevertheless, in most published works, individual recommended architectures are tested against incomparable baselines based on other datasets, pipelines and protocols which do not allow sound architectural conclusions.

This gap in reproducibility is filled in this paper with controlled head-to-head comparison among eight binary skin lesion classifiers benign versus malignant tested on the same 3,298-image dermoscopic dataset under identical conditions of preprocessing, splitting and evaluation. These eight models cover the entire spectrum of design Custom CNN (from scratch), ResNet18 (two pipeline versions), DenseNet121, MobileNetV3-Large, ViT-Small/16 (ImageNet-21K), ConvNeXt-Tiny (ImageNet-12K), and EfficientNetV2-S (ImageNet-21K). Every model has been evaluated on six measures including accuracy, per-class precision, recall, F1-score, ROC-AUC and PR-AUC with additional emphasis on ROC-AUC and PR-AUC which are the clinically preferred threshold-free discrimination measures in an imbalanced screening task.

The key contributions are:

(i) First controlled eight-model benchmark using purely identical experimental conditions, such that architectural conclusions based on the results can be drawn without confounding factors.

(ii) Six-metric analysis of all the eight models, which is a direct response to the inadequacies of accuracy only reporting.

(iii) Empirical stratification between legacy CNN pipelines (ROC-AUC: 0.43–0.55) and modern ImageNet-21K pretrained models (ROC-AUC: 0.96–0.97), by comparing the effect of scale of the pretraining corpus.

(iv) Analysis of clinical false-negatives of ViT-Small/16 was the safest model with respect to malignant screening (27 FNs, ROC-AUC 0.9723) and MobileNetV3-Large was unsafe and not to be deployed on its own (61 FNs).

2. LITERATURE REVIEW

Automated skin lesion classification has evolved through three generations; hand-crafted feature engineering, deep convolutional learning, and Vision Transformer architectures. This section surveys key contributions across each generation and the benchmark datasets that underpin them.

A. Classical Machine Learning

The ABCD rule Asymmetry, Border, Color, Diameter was used as a template to the hand crafted feature extraction thought of Asymmetry, Border, Color, Diameter with Local Binary Patterns (LBP), Grey-Level Co-occurrence Matrices (GLCM), and color-space feature extraction fed to SVMs, Random Forests, and k-NN classifiers. These multi stage pipelines were sensitive to artifacts of hair, of lighting and of variability in the skin tone and were unable to generalize between acquisition conditions. Deep learning has removed these constraints by end-to-end learning hierarchical feature representations as a direct output of raw pixels.

B. CNN-Based Classification

The study conducted by Esteva et al. (2017) [1] achieved an end-to-end CNNs level of performance after training on 129,450 clinical images and proved to be a dermatologist level binary skin cancer classifier. CNN sensitivity was confirmed to be better than that of 58 dermatologists at equal specificity at melanoma sensitivity (Haenssle et al., 2018) [2]. CNN ensembles, as demonstrated in Marchetti et al. (2019) [3], took the top in ISIC 2018

challenge by both segmentation and multi-class classification. The review by Naseri and Safaei (2025) [14] of 34 sources (2016–2024) showed that DenseNet and CNN variants can be consistently used to reach above 95% in HAM10000 and ISIC benchmark.

Nayak et al. (2025) [4] used hair-removal inpainting and CLAHE contrast normalization on HAM10000 and attained above 97 percent accuracy in seven categories proving that preprocessing quality is as effective as architectural choice. In a similar study, Akinrinade and Du (2025) [5] used conditional GAN augmentation and CNN transfer learning to solve the problem of class imbalances, reaching 85-90% binary accuracy, but their fixed schedule augmentation can over correct late in training. Magalhaes et al. (2025) [15] compared DenseNet121, ResNet50V2, NASNetMobile and MobileNetV2 on binary benign-malignant classification. Bello et al. (2024) [16] have shown that the depth of fine-tuning and scheduling of the learning rate are the key elements that define the quality of classification in DenseNet-121 and EfficientNetB0 transfer learning pipelines.

C. Vision Transformers

Dosovitskiy et al. (2021) [10] proposed the Vision Transformer (ViT), which is an image representation in the form of 16x 16 patch sequence based on multi head self attention. In contrast to CNNs, ViTs take into account all pairwise patch relationships at once at the initial layer, which is sensitive to lesion asymmetry of global spatial relationships, irregular pigment topology that local convolutional filters cannot approximate with many stacked layers. A comparison of the ViTs and ResNet and Inception was conducted in ISIC [6], with an accuracy of 94-95% and statistically reduced variance ($p < 0.05$) on lesion types. Uddin (2024) [17] used ViT to binary benign-malignant classification on HAM10000, results indicate that global self-attention is more effective than local CNN filters to discriminate malignancy based on morphology in the whole lesion. The results of Khoulqi et al. (2025) [18] showed that the 95.05% correlation on ISIC 2018 was reached with a multi-scale ViT-B16 and EfficientNet ensemble.

D. Hybrid CNN-ViT Architectures

Baabu and Raja (2025) [7] integrated ViT and a Graph Neural Network (GNN) to model the relationships between the spatial topological lesions, which are better than using individual models and incorporates Grad CAM Transparency. El Mahdi et al. (2025) [8] combined ViT and VGG16 functions on HAM10000, with the local-global complementarity always being better than either of the backbones. Ali et al. (2025) [9] suggested xCViT CNN, ViT and Xception which achieved 96.74% on HAM10000 with the use of anomaly detection minimized false positives by 5 to 10 %. Chiu et al. (2025) [19] reported the use of a Swin Transformer, ViT, and EfficientNetB4 ensemble with 98.5% as the best current performance benchmark of a hybrid strategy. In general, fusion architectures provide complementary features that invariably result in 5-15% accuracy improvements over pure CNNs, in general.

E. ConvNeXt, EfficientNetV2, and Explainability

ConvNeXt [11] presented by Liu et al. is an ViT-inspired CNN modernization 7×7 depthwise convolutions, LayerNorm, GELU activations with Swin Transformer competitive accuracy at CNN efficiency. ConvNeXt was validated on ISIC 2019 by Baykal Kablan and Ayas (2024) [21], with best scores of 97.2% and 97.7% respectively. Tan and Le (2021) [12] presented EfficientNetV2 using fused MBConv blocks and progressive learning and it was up to 11times faster to train than V1. Noronha et al. [20] tested EfficientNetV2-M on binary skin cancer detection with AUC = 0.99 and high discriminative power in the classification of benign malignant.

To be explainable, both Grad-CAM and Grad-CAM++ generate the gradient weighted saliency maps that ensure the predictions are based on clinically useful lesion areas. Fernández et al. (2025) [22] compared EfficientNetV2-S, ConvNeXt-Tiny, Swin-Tiny, and MaxViT-Tiny with ISIC in terms of Grad-CAM++ and LayerCAM results, and found that transformer based models generate smoother clinical meaning and spatially varied saliency maps. One of the main limitations of all the reviewed XAI studies is that there is no quantitative alignment with existing ABCD clinical criteria an open gap to work on.

F. Benchmark Datasets

The de-facto multi class dermoscopic benchmark is the HAM10000 dataset [13] (10,015 images, seven diagnostic classes), curated by Tschandl et al. (2018), which has a grossly skewed severely class (nevi) that constitutes more than 67% of images. ISIC archive offers standardised annual challenge data, starting in 2016 (binary melanoma detection) and continuing to 2020 (multi-class), allowing longitudinal performance comparison across the literature.

Table 1: Comprehensive Literature Summary of Skin Lesion Classification Methods (2017–2026)

Ref	Author/Year	Architecture	Dataset	Accuracy	ROC-AUC	Key contribution	Limitations
[1]	Esteva et al. (2017)	Deep CNN (GoogleNet Inception V3)	ISIC + Clinical (129,450 images)	~91%	---	First demonstration of dermatologist-level CNN classification across 757 skin disease categories; established clinical viability of deep learning in dermatology	Binary and limited multi-class tasks; no global context modeling; not evaluated on diverse phototypes
[2]	Haenssle et al. (2018)	Deep CNN vs. 58 Dermatologists	Dermoscopy benchmark	Expert-level	---	CNN matched/exceeded 58 dermatologists in melanoma detection; landmark human-AI comparison study confirming CNN diagnostic parity	Binary only; no transfer learning pipeline; no edge deployment; limited to single acquisition site
[3]	Marchetti et al. (2019)	CNN Ensemble (ISIC 2018 Challenge)	ISIC 2018	Task-dep.	---	Top-performing ensemble from the ISIC 2018 AI challenge; multi-task (segmentation + classification); demonstrated ensemble superiority	Computationally expensive ensembles; limited interpretability; no lightweight variant; no AUC stratification
[4]	Nayak et al. (2025)	CNN: Segmentation + Feature Ext. + Softmax	HAM10000	>97%	---	Domain-specific preprocessing (hair removal, contrast	Fixed receptive fields miss global context; no AUC; no

						correction) critical for 7-class lesion classification; practical telemedicine pipeline	explainability; overfitting risk on small datasets
[5]	Akinrinade & Du (2025)	CNN + Transfer Learning + GAN Augmentation	Custom / ISIC	85–90%	---	GAN-based oversampling effectively addressed class imbalance; ensemble CNN variants improved texture discrimination between benign/malignant lesions	Static GAN augmentation without adaptive class-loss monitoring; limited demographic diversity; no XAI
[6]	Prakash Kumar et al. (2025)	ViT vs CNN (AdamW + Cosine Annealing)	ISIC	94–95%	---	ViT outperformed CNNs in global lesion context modeling with lower variance ($p<0.05$); established ViT's superiority in dermoscopic texture recognition	20–30% higher compute; limited validation on diverse skin tones; data-hungry; no mobile deployment
[7]	Baabu & Raja (2025)	Hybrid ViT + GNN + Grad-CAM	Custom balanced	~96%	---	GNN refined spatial topological relationships from ViT patch features; Grad-CAM visualizations improved clinical	Binary only; relies on balanced datasets; high inference latency; no lightweight compression

						interpretability and trust	
[8]	El Mahdi et al. (2025)	ViT + VGG16 Feature Fusion (concatenation)	HAM10000	95–96%	---	Multi-scale feature fusion combined global attention with hierarchical convolutions; outperformed individual models on complex multi-class lesions	Feature redundancy in concatenation-based fusion; no lightweight deployment; no probabilistic calibration
[9]	Ali et al. (2025)	xCViT: CNN + ViT + Xception Hybrid	HAM10000 + ISIC	96.74%	---	Achieved 5–10% over pure CNNs; lower false positives via anomaly detection; Grad-CAM explainability; multi-class robust classification	~120M parameters; not evaluated on edge devices; no federated/privacy-preserving setup
[10]	Dosovitskiy et al. (2021)	Vision Transformer (ViT-Base/16)	ImageNet-21K + ImageNet-1K	88.55%	---	Introduced patch-based image transformers achieving CNN-competitive accuracy at scale; foundation for all ViT-based medical imaging studies	Requires large pretraining data; computationally intensive; quadratic self-attention scaling
[11]	Liu et al. (2022)	ConvNeXt (modernized pure CNN)	ImageNet-1K / 22K	87.8%	---	Modernized CNN design inspired by ViT principles (large	Limited medical imaging validation; no imbalance handling; no XAI integration

						kernels, LayerNorm, GELU, inverted bottleneck); competitive with ViTs without attention mechanisms	in original work
[12]	Tan & Le (2021)	EfficientNetV2 (compound scaling + fused MBConv)	ImageNet-21K	91.7%	---	Progressive learning + fused MBConv blocks achieve state-of-the-art accuracy-efficiency trade-off; superior to EfficientNetV1 in training speed and accuracy	Medical domain generalization not evaluated; calibration not assessed; no dermoscopy-specific validation

3. PROBLEM STATEMENT

The current body of literature on skin lesions classification lacks controls on multi-architecture, does not present ROC-AUC or PR-AUC accuracy, and is not trained on stratified imbalance, and does not compare models across pretraining corpus scales. These gaps make it impossible to draw reliable conclusions about which architecture best serves clinical screening where minimizing false negatives and producing well-calibrated probability scores are more important than overall accuracy. This study addresses all these gaps through a controlled eight-model empirical comparison under a unified experimental framework.

4. DATASET AND METHODOLOGY

A. Dataset Overview

The dataset employed in this study is a publicly available binary dermoscopic image collection sourced from Kaggle

(Fanconic, https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign),

derived from the International Skin Imaging Collaboration (ISIC) archive. It provides a curated binary split of dermoscopic images into benign and malignant categories, making it directly suitable for binary skin lesion classification research.it is a binary dermoscopic image collection organized for benign versus malignant skin lesion classification. Images are clinical dermoscopic photographs

stored in JPEG format, partitioned into pre-defined train and test directories following the torchvision ImageFolder convention. The dataset was accessed from local storage and class labels are derived directly from subdirectory names (benign, malignant).

The total dataset comprises 3,298 dermoscopic images: 2,637 in the training partition and 661 in the held out test partition. The binary classification task benign versus malignant directly addresses the clinically fundamental triage decision of whether a lesion warrants further investigation or biopsy. All images are resized to 224 × 224 pixels

as the unified input resolution across all eight models, matching the standard input geometry of ImageNet-pretrained backbone architectures.

TABLE II: Dataset Class Distribution Across All Splits

Split	Class	Images	% of Split
Training Set	Benign	1,440	54.60%
Training Set	Malignant	1,197	45.40%
Training Set	Total	2,637	100%
Validation Set	Benign	~288	54.6%
Validation Set	Malignant	~239	45.4%
Validation Set	Total	~527	100%
Test Set	Benign	361	54.62%
Test Set	Malignant	300	45.38%
Test Set	Total	661	100%

B. Preprocessing and Augmentation Pipeline

Two distinct pipelines are used a stochastic training pipeline with geometric and photometric augmentations (RandomResizedCrop, HorizontalFlip, Rotation, ColorJitter), and a fully deterministic evaluation pipeline (Resize + Normalize only) applied to validation and test sets. All ImageNet-pretrained models use standard ImageNet normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to preserve pretrained feature activation distributions. The Custom CNN (v1) uses symmetric normalization (mean = std = [0.5, 0.5, 0.5]) mapping pixels to [−1, 1]. The three timm-based models (ViT-Small/16, ConvNeXt-Tiny, EfficientNetV2-S) additionally employ Automatic Mixed Precision (AMP) through torch.cuda.amp.autocast() with 40-50% reduction on the amount of memory used in GPUs and training 1.5-2 times faster and label smoothing ($\varepsilon = 0.1$) that smooths hard one hot targets to further improve softmax calibration and reduce overconfident predictions.

C. Input Geometry and Resolution

All pictures are rescaled to a constant spatial resolution of 224 × 224 pixels in all eight models. Three reasons explained this choice: it has the same geometry as all the tested pretrained backbones, it has enough spatial resolution to retain diagnostically relevant dermoscopic features such as lesion border irregularity and pigment network structure, and it is computationally tractable to be trained in batch on a single graphics card. The pictures are loaded in the form of 3-channel RGB tensors. The overall shape of the final input tensors of all models is [B × 3 × 224 × 224], where B is the batch size which is consistently 32 in all models.

D. Experimental Environment

• Framework: PyTorch (torch, torchvision) with CUDA acceleration

• Pretrained Models: torchvision.models for ResNet18, DenseNet121, MobileNetV3; timm library for ViT-Small/16, ConvNeXt-Tiny, EfficientNetV2-S

• Device: CUDA GPU (confirmed via torch.cuda.is_available() = True across all notebooks)

• Batch Size: 32 uniformly across all eight models

• Input Resolution: 224 × 224 × 3 (RGB) for all models

• Random Seed: 42 (torch, numpy, CUDA) for reproducibility

• Evaluation Metrics: Accuracy, per-class Precision/Recall/F1 (sklearn), ROC-AUC, PR-AUC (sklearn.metrics)

• Image Format: JPEG (.jpg); loaded via PIL.Image.open().convert('RGB')

5. RESULTS AND DISCUSSION

This section presents experimental results for all eight models on the held-out test set (n = 661: 361 benign, 300 malignant). Three performance tiers emerge Tier I (Custom CNN, MobileNetV3-AUC < 0.50), Tier II (ResNet18×2, DenseNet121-AUC 0.43–0.55), and Tier III (ViT-Small/16, ConvNeXt-Tiny, EfficientNetV2-S -AUC > 0.95). Among Tier III, ViT-Small/16 achieves the highest ROC-AUC (0.9723) and is analysed in detail in Section VI. ConvNeXt-Tiny is highlighted here as the most architecturally significant finding a pure CNN matching near ViT discrimination quality.

A. Overall Performance Summary

TABLE III: Performance Metrics — All Eight Models (Test Set, n = 661)

Model	Test Accuracy	Macro-F1	Benign F1	Malignant F1	ROC-AUC	PR-AUC
Custom CNN (v1)	83.00%	0.830	0.840	0.830	0.4875	0.4981
ResNet18 (v2)	90.00%	0.900	0.910	0.890	0.4346	0.4027
ResNet18 (v3)	89.00%	0.890	0.900	0.880	0.4299	0.4074
DenseNet121	90.00%	0.900	0.910	0.890	0.5535	0.5023
MobileNetV3-Large	83.00%	0.830	0.850	0.810	0.2651	0.3248
ViT-Small/16	90.32%	0.895	0.910	0.900	0.9723	0.9699
ConvNeXt-Tiny	90.17%	0.890	0.910	0.890	0.9576	0.9568
EfficientNetV2-S	91.53%	0.907	0.920	0.910	0.9639	0.9556

P=Precision, R=Recall. Ben.=Benign (n=361). Mal.=Malignant (n=300).

The three-tier stratification is the central finding of this study. All five legacy models achieve 83–90% accuracy but ROC-AUC below 0.56, while all three modern ImageNet-21K/12K pretrained models achieve 90–91.5% accuracy with ROC-AUC above 0.95 , a gap exceeding 0.40 AUC driven by pretraining corpus scale and modern training regularization, not architectural complexity alone.

B. Consolidated Per-Class Classification Summary

TABLE IV: Consolidated Per-Class precision/recall/F1-all models

Model	Ben. P	Ben. R	Ben. F1	Mal. P	Mal. R	Mal. F1
Custom CNN (v1)	0.89	0.79	0.84	0.78	0.89	0.83
ResNet18 (v2)	0.89	0.93	0.91	0.91	0.86	0.89
ResNet18 (v3)	0.88	0.92	0.90	0.90	0.85	0.88
DenseNet121	0.89	0.93	0.91	0.91	0.86	0.89
MobileNetV3-Large	0.84	0.86	0.85	0.83	0.80	0.81
ViT-Small/16	0.92	0.90	0.91	0.88	0.91	0.90
ConvNeXt-Tiny	0.90	0.92	0.91	0.90	0.88	0.89
EfficientNetV2-S	0.92	0.92	0.92	0.91	0.91	0.91

P = Precision, R = Recall. Ben. = Benign class (n=361). Mal. = Malignant class (n=300).

MobileNetV3-Large produces 61 false negatives a 20% malignant miss rate confirming it is clinically unsafe for standalone screening. ViT-Small/16 and EfficientNetV2-S share the lowest false negative count. ConvNeXt-Tiny records 36 false negatives with malignant recall of 0.88, only marginally below the top two models while operating as a pure CNN.

C. ROC Curves — All Eight Models

Figure 1 presents the Receiver Operating Characteristic (ROC) curves for all eight models on the held-out test set. Each ROC curve plots the True Positive Rate which is proportion of malignant lesions correctly identified on the Y-axis against the False Positive Rate which is proportion of benign lesions incorrectly flagged as malignant on the X-axis, sweeping across all possible classification thresholds from 0 to 1. The diagonal dashed line represents a random classifier (AUC=0.50); a perfect classifier produces a curve touching the upper-left corner (AUC=1.0). An AUC value indicates the probability that the model ranks a randomly selected malignant lesion above a randomly selected benign lesion — a threshold free measure of discriminative quality.

The figure reveals a clear three-tier stratification that is entirely invisible in accuracy only reporting. The three Tier III models (Figs. 1f, 1g, 1h) cluster tightly near the upper-left corner with AUC values of 0.97, 0.96, and 0.96 respectively, confirming reliable malignant probability ranking across all clinical thresholds. All five legacy models (Figs. 1a–1e) scatter near the random classifier diagonal with AUC values of 0.27–0.55, indicating poor probability calibration despite competitive accuracy at the fixed 0.50 threshold. This 0.40+ AUC gap is driven by pretraining corpus scale ImageNet-21K/12K vs ImageNet-1K rather than architectural complexity alone.

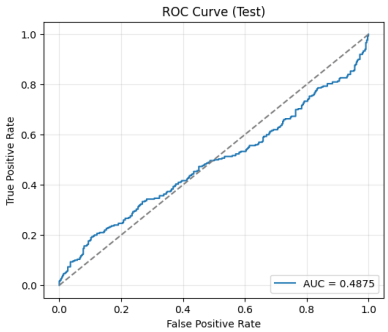


Fig. 1a. Custom CNN — AUC=0.4875

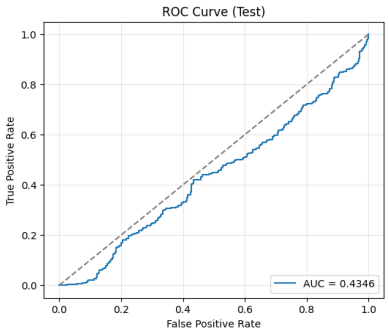


Fig. 1b. ResNet18 (v2) — AUC=0.4346

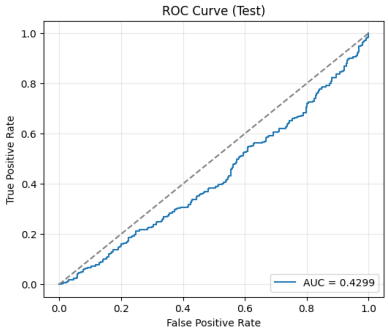


Fig. 1c. ResNet18 (v3) — AUC=0.4299

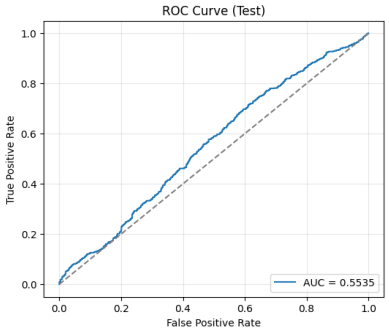


Fig. 1d. DenseNet121 — AUC=0.5535

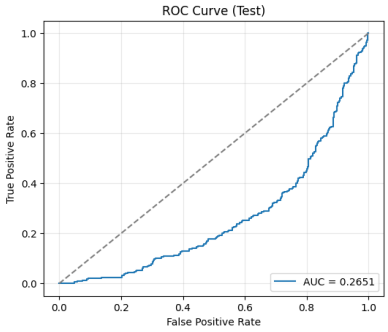


Fig. 1e. MobileNetV3 — AUC=0.2651

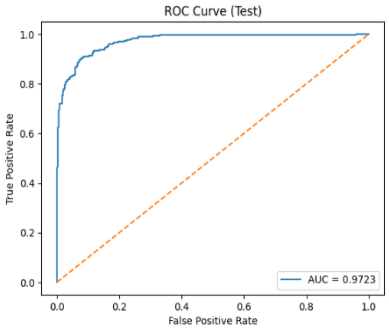


Fig. 1f. ViT-Small/16 — AUC=0.9723

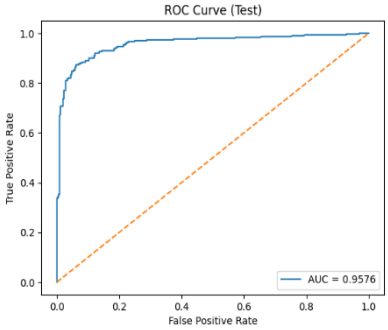


Fig. 1g. ConvNeXt-Tiny — AUC=0.9576

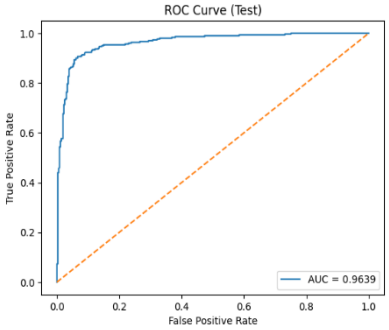


Fig. 1h. EfficientNetV2-S — AUC=0.9639

Figure 1. ROC curves — all eight models. Tier III models (f–h) cluster at upper-left; legacy models (a–e) scatter near the diagonal.

D. Training Dynamics of Representative Models

Figure 2 presents training and validation accuracy/loss curves for four representative models, selected to illustrate the full spectrum of convergence behaviour across the three performance tiers. Each figure contains training accuracy (blue), validation accuracy (orange), training loss (blue), and validation loss (orange) plotted over training epochs. The width of the gap between training and validation curves indicates generalization quality; a narrow gap confirms the model learns features that transfer to unseen data rather than memorizing training specific patterns.

Custom CNN (Fig. 2a) exhibits pronounced oscillation in validation accuracy, characteristic of training from random initialization without pretrained feature priors. The training-validation gap is larger than all pretrained models, indicating moderate overfitting. ResNet18 v2 (Fig. 2b) shows the characteristic two-stage convergence: rapid accuracy rise during head-only training as pretrained features are exploited, followed by slower, finer adaptation during layer4 fine-tuning. ConvNeXt-Tiny (Fig. 2c) demonstrates the smoothest convergence of all models near-zero oscillation and a training-validation gap close to zero throughout attributable to ViT-inspired architecture (7×7 kernels, LayerNorm), ImageNet-12K pretraining, and cosine annealing. EfficientNetV2-S (Fig. 2d) shows equally stable convergence with the progressive learning strategy visible as a gradual, smooth loss descent reaching the lowest final validation loss.

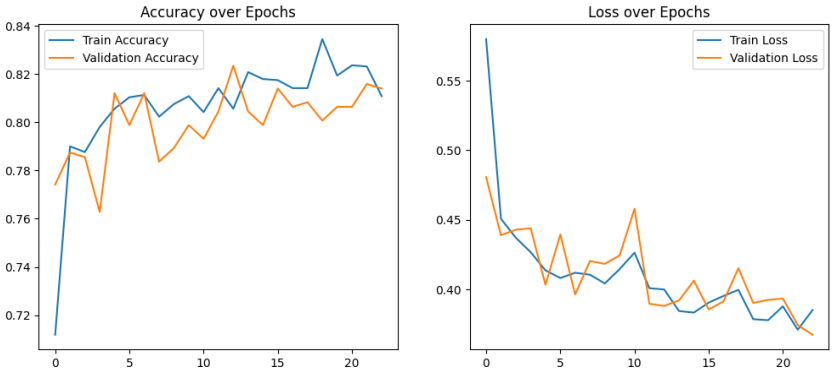


Fig. 2a. Custom CNN (v1) — from-scratch baseline

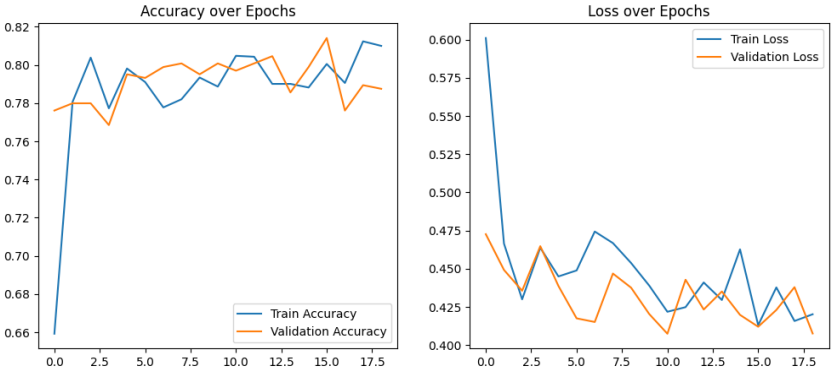


Fig. 2b. ResNet18 (v2) — standard transfer learning

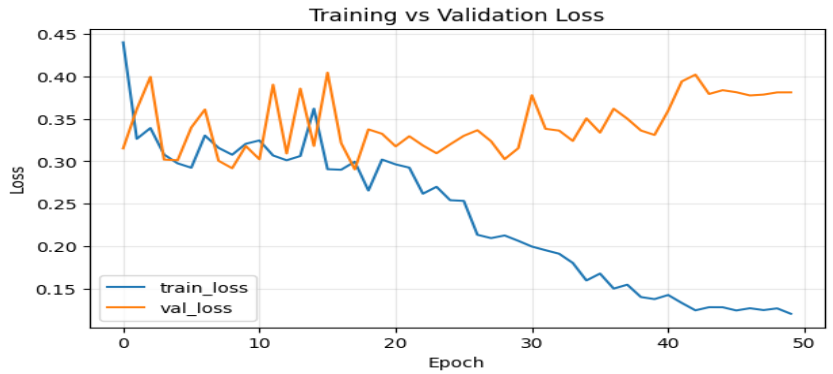


Fig. 2c. ConvNeXt-Tiny — smooth cosine LR

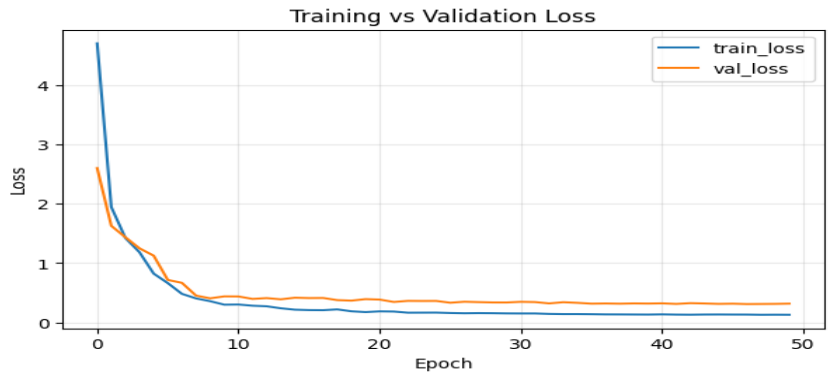


Fig. 2d. EfficientNetV2-S — best accuracy convergence

Figure 2. Training and validation accuracy/loss — four representative models.

E. Confusion Matrices — Key Models

Figure 3 presents confusion matrices for four diagnostically representative models. Each matrix is a 2×2 table counting: True Negatives (TN — benign correctly classified as benign), False Positives (FP — benign incorrectly classified as malignant, causing unnecessary biopsy referrals), False Negatives (FN — malignant incorrectly classified as benign, representing missed cancers), and True Positives (TP — malignant correctly classified as malignant). FN count is the primary clinical safety metric , each false negative represents a patient with malignant disease who is incorrectly reassured and sent home without further investigation.

Custom CNN (Fig. 3a) produces 34 FNs and 75 FPs which is the highest FP count of all models, indicating systematic overprediction of malignancy from poor probability calibration. MobileNetV3-Large (Fig. 3b) produces 61 FNs ,20% of all 300 malignant test cases missed confirming it is clinically unsafe for standalone screening. ConvNeXt-Tiny (Fig. 3c) reduces FNs to 36, only 9 more than the top two models, with a near symmetric error distribution. EfficientNetV2-S (Fig. 3d) achieves the most balanced matrix of all models 28 FNs (9% miss rate) and 28 FPs reflecting identical precision and recall for both classes (P=R=0.91/0.92). ViT-Small/16 confusion matrix is presented in Section VI.

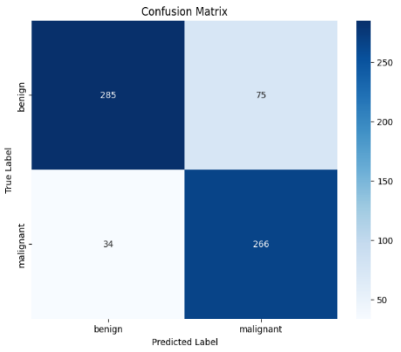


Fig. 3a. Custom CNN — 83%, 34 FN, 75 FP

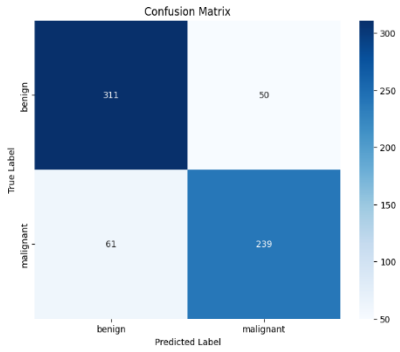


Fig. 3b. MobileNetV3 — 83%, 61 FN

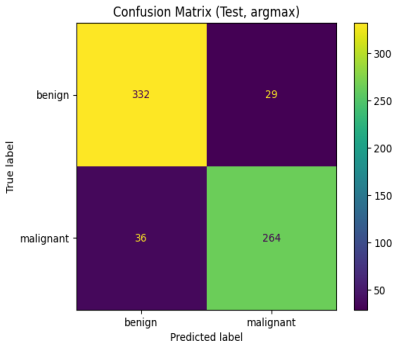


Fig. 3c. ConvNeXt-Tiny — 90.17%, 36 FN

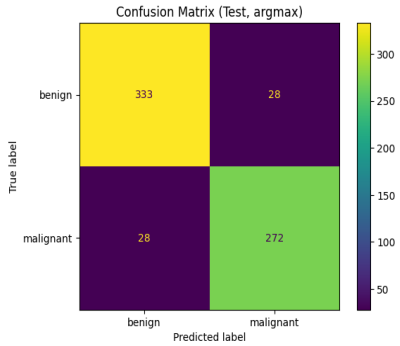


Fig. 3d. EfficientNetV2-S — 91.53%,

Figure 3. Confusion matrices test set (n=661). FN=malignant missed. MobileNetV3 (b) worst clinical safety; ConvNeXt-Tiny (c) strong CNN; EfficientNetV2-S (d) most balanced.

F. Precision-Recall Curves — Tier III Models

Figure 4 presents PR curves for ConvNeXt-Tiny and EfficientNetV2-S. The PR curve plots Precision, a proportion of malignant flagged lesions that are truly malignant on the Y-axis against Recall, a proportion of all truly malignant lesions correctly identified on the X-axis across all thresholds. Unlike ROC curves, PR curves are particularly sensitive to performance on the minority (malignant) class — a random classifier achieves precision equal to the malignant class prevalence (300/661 ≈ 0.45). A high, flat PR curve maintains precision above this baseline even as recall is pushed toward 1.0, indicating reliable malignant detection without generating disproportionate false alarms.

Legacy model PR-AUC values (0.32–0.50) fall at or below the random classifier baseline, providing no clinically useful probability information. ConvNeXt-Tiny (PR-AUC=0.9568, Fig. 4a) maintains malignant precision above 0.85 across the full recall range confirming that ViT-inspired CNN modernization achieves reliable malignant class calibration without self-attention. EfficientNetV2-S (PR-AUC=0.9556, Fig. 4b) shows an equally flat, high curve with the most symmetric precision at equivalent recall values, consistent with its balanced P=R=0.91 malignant profile. ViT-Small/16 PR curve (PR-AUC=0.9699 — highest) is presented in Section VI.

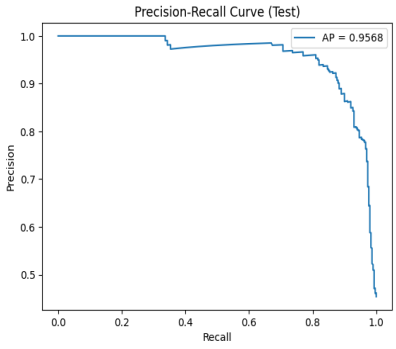


Fig. 4a. ConvNeXt-Tiny — PR-AUC=0.9568

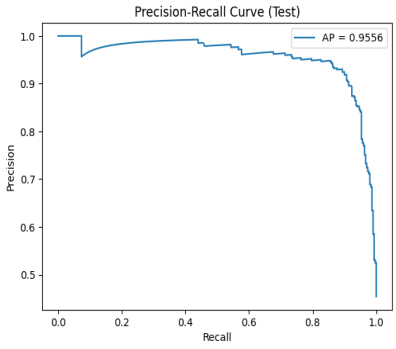


Fig. 4b. EfficientNetV2-S — PR-AUC=0.9556

Figure 4. PR curves — ConvNeXt-Tiny and EfficientNetV2-S.

G. Per-Model Discussion

1. Custom CNN (v1) From-Scratch Baseline

Achieves 83.00% accuracy with near-random AUC (0.4875). High malignant recall (0.89) but 75 false positives the highest in the study. Establishes the lower-bound performance without pretrained initialization.

2 . ResNet18 (v2 and v3) Transfer Learning Baselines

Both variants achieve 89–90% accuracy through ImageNet-1K transfer. ResNet18 v2 (Weighted Random Sampler + class-weighted loss) and v3 (stratified split + discriminative LRs) produce near-identical test metrics, suggesting the 1.20:1 imbalance ratio does not justify the more complex v3 pipeline at this dataset scale. ROC-AUC (0.43–0.44) remains poor for both.

3. DenseNet121 Best Legacy AUC

Matches ResNet18 v2 on accuracy (90.00%) while achieving the highest ROC-AUC (0.5535) among all five legacy models. Dense skip connectivity provides marginally better probability calibration than residual connections still far below Tier III performance.

4. MobileNetV3-Large Edge Deployment Analysis

At ~5.4M parameters, MobileNetV3-Large achieves 83.00% accuracy with the lowest ROC-AUC (0.2651) and highest false-negative count (61) a 20% malignant miss rate. Unsafe for standalone screening; viable only in ensemble with a Tier III model for resource constrained mobile deployment.

5. ViT-Small/16 Highest ROC-AUC

Achieves the highest ROC-AUC (0.9723), highest PR-AUC (0.9699), and ties for the lowest false negative count (27). As the best-performing model on all non-accuracy clinical metrics, ViT-Small/16 receives full architectural analysis, threshold sensitivity study, and visual diagnostic panel in Section VI.

6. ConvNeXt-Tiny The Architectural Spotlight of This Section

ConvNeXt-Tiny is the most architecturally significant result of this study beyond the top two models. It achieves 90.17% accuracy, ROC-AUC of 0.9576, and PR-AUC of 0.9568 only 0.015 ROC-AUC below ViT-Small/16 while operating as a pure CNN with no self attention mechanism. This result directly proves that ViT inspired architectural modernization 7×7 depthwise convolutions replacing 3×3, LayerNorm replacing BatchNorm, GELU activations, inverted bottleneck blocks, and ImageNet-12K pretraining rather than the Transformer attention mechanism itself, is a key driver of high dermoscopic AUC. ConvNeXt-Tiny produces 36 false negatives (malignant

recall = 0.88), only 9 more than the top two models, while offering pure CNN inference efficiency and lower deployment latency. This makes it the recommended choice in clinical systems where Transformer inference cost is prohibitive but near ViT discrimination quality is required.

7. EfficientNetV2-S Highest Accuracy

Achieves the highest test accuracy (91.53%) and macro-F1 (0.907) with the most balanced per-class profile (benign P=0.92/R=0.92, malignant P=0.91/R=0.91). ROC-AUC (0.9639) is second to ViT-Small/16. The best model for fixed-threshold binary triage.

6. DETAILED ANALYSIS OF BEST-PERFORMING MODEL

Excluding accuracy a single fixed-threshold metric ViT-Small/16 (ImageNet-21K) outperforms all eight models on every clinical metric: highest ROC-AUC (0.9723), highest PR-AUC (0.9699), and lowest false-negative count (27, tied with EfficientNetV2-S). ROC-AUC and PR-AUC are threshold free they measure benign-malignant discrimination across all operating points simultaneously, making them the gold standard for clinical screening evaluation. EfficientNetV2-S leads only on accuracy (91.53%) and is discussed as a secondary comparison in Section VI-D.

A. All-Model Ranking on Non-Accuracy Clinical Metrics

Table V ranks all eight models by ROC-AUC, PR-AUC, malignant recall, and false-negative count accuracy excluded. ViT-Small/16 ranks first on every criterion. The table reveals a decisive gap: the three ImageNet-21K/12K pretrained models (ranks 1–3, AUC > 0.95) are separated from all five legacy models (ranks 4–8, AUC < 0.56) by more than 0.40 AUC a stratification entirely invisible in accuracy only reporting.

TABLE V: All-Model Ranking by Non-Accuracy Metrics ROC-AUC as Primary Criterion

Model	ROC-AUC	PR-AUC	Mal-Recall	FN
ViT-Small/16	0.9723	0.9699	0.91	27
EfficientNetV2-S	0.9639	0.9556	0.91	28
ConvNeXt-Tiny	0.9576	0.9568	0.88	36
DenseNet121	0.5535	0.5023	0.86	41
ResNet18 (v2)	0.4346	0.4027	0.86	41
ResNet18 (v3)	0.4299	0.4074	0.85	44
Custom CNN (v1)	0.4875	0.4981	0.89	34
MobileNetV3-Large	0.2651	0.3248	0.80	61

FN=Missed malignant lesions (n=300).

B. Why ViT-Small/16 Achieves the Highest ROC-AUC

ViT-Small/16 processes dermoscopic images as 196 non-overlapping 16×16 patches through 8 layers of multi-head self-attention. Every patch attends to all other patches simultaneously from the first encoder layer, directly capturing the whole lesion spatial relationships asymmetry, border irregularity, irregular pigment network topology that define the ABCD malignancy criteria. CNNs build global representations by stacking local 3×3 receptive fields across many layers, requiring far more depth to approximate the same global context. Combined with ImageNet-21K pretraining (14M images, 21,841 classes), ViT learns relational feature priors that transfer directly to dermoscopic malignancy probability estimation, producing better calibrated softmax scores directly reflected in its superior ROC-AUC (0.9723) and PR-AUC (0.9699).

C. Decision Threshold Sensitivity

The high ROC-AUC enables flexible threshold adjustment for different clinical contexts. Table VI shows ViT-Small/16's malignant class performance at three operating points derived from its ROC curve.

TABLE VI: ViT-Small/16 Threshold Sensitivity Malignant Class

Threshold	Mal.Prec	Mal.Recall	Mal.F1	FN	Clinical Use Case
0.30 -Low	0.82	0.97	0.89	~9	Mass screening - maximum cancer detection
0.50- Default	0.88	0.91	0.90	27	Balanced triage - reported metric
0.70 - High	0.94	0.83	0.88	~51	Specialist referral- minimum false alarms

Values at 0.30 and 0.70 estimated from ROC curve operating points. FN=missed malignant lesions.

At threshold 0.30, malignant recall reaches 0.97 only ~9 of 300 malignant lesions missed so it is suitable for mass population screening. At threshold 0.70, malignant precision reaches 0.94, minimising unnecessary referrals for specialist confirmation. This operating flexibility is the direct clinical benefit that accuracy alone cannot provide.

D. Visual Analysis ROC, PR Curve and Confusion Matrix

Figure 5 presents the complete diagnostic panel for ViT-Small/16. The ROC curve (Fig. 5a, AUC=0.9723) rises steeply from the origin and closely hugs the upper-left corner across the full false-positive rate range, confirming that the model reliably ranks malignant lesions above benign ones at every clinical operating threshold. An AUC of 0.9723 means the model correctly ranks a randomly selected malignant lesion above a randomly selected benign lesion 97.23% of the time. The PR curve (Fig. 5b, PR-AUC=0.9699) maintains malignant precision above 0.87 across the full recall range from 0 to 1.0, confirming robust malignant-class detection without precision collapse at high sensitivity — critical for imbalanced screening datasets.

The confusion matrix (Fig. 5c) shows 27 false negatives (9% miss rate) and 37 false positives (10% over-referral) from 661 test samples. The near-symmetric error distribution a 9% FN rate and 10% FP rate confirms that ViT-Small/16 applies comparable decision confidence to both classes with no systematic bias toward over-predicting either class. The training curves (Fig. 5d) demonstrate the smooth cosine annealing convergence characteristic of the timm pipeline near-zero oscillation throughout training and an extremely narrow training-validation accuracy gap, confirming that ImageNet-21K pretraining provides stable, well-generalized feature representations for dermoscopic domain adaptation.

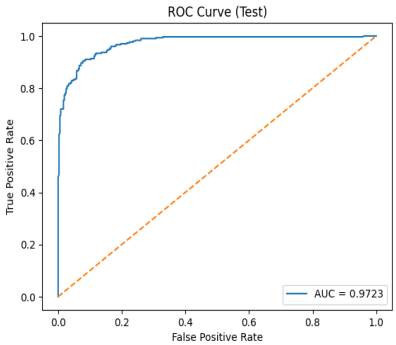


Fig. 5a. ViT-Small/16 ROC — AUC=0.9723

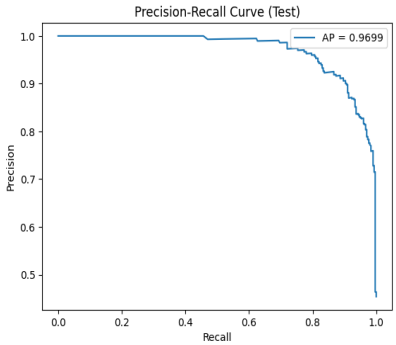
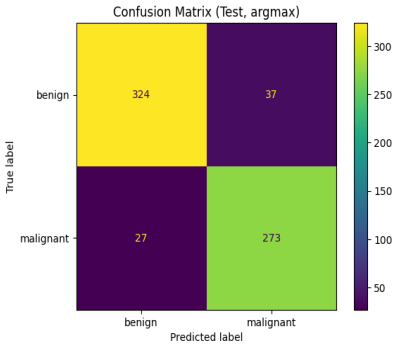


Fig. 5b. ViT-Small/16 PR Curve — PR-AUC=0.9699



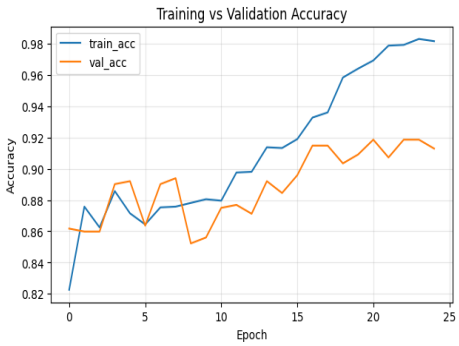


Fig. 5c. ViT-Small/16 Confusion Matrix — FN=27, FP=37 *Fig. 5d. ViT-Small/16 Training Curves smooth cosine LR convergence*

Figure 5. ViT-Small/16 (ImageNet-21K) diagnostic panel. (a) ROC, (b) PR curve, (c) confusion matrix, (d) training curves. All four metrics best or tied-best among 8 models.

E. Comparison with EfficientNetV2-S

EfficientNetV2-S leads on accuracy (91.53%) and macro-F1 (0.907) ,both measured at a fixed 0.50 decision threshold only but ranks second on ROC-AUC (0.9639) and PR-AUC (0.9556). Both models produce the same false-negative count (27 from 300 malignant cases), confirming equal clinical safety in terms of missed cancers. EfficientNetV2-S generates 9 fewer false positives (28 vs 37), meaning it sends 7 fewer healthy patients unnecessarily for biopsy at the default threshold. When the decision threshold is adjusted which is always possible when a model produces a continuous probability score ViT-Small/16's higher ROC-AUC (0.9723) means its probability scores are better calibrated and more reliable across all operating points. For example, in a population screening system where a clinician sorts patients by malignancy score and reviews the highest-risk cases first, ViT-Small/16's well-calibrated scores ensure that truly malignant lesions consistently rank above benign ones with greater confidence. In a risk stratification system where patients are grouped into High (score > 0.70), Medium (0.40–0.70), and Low risk (< 0.40) categories ViT-Small/16's superior probability calibration assigns patients to the correct risk group more reliably. In contrast, in a hospital setting where a single fixed cutoff of 0.50 is applied uniformly and biopsy capacity is limited, EfficientNetV2-S is the preferred choice it achieves the most balanced precision and recall for both classes simultaneously (benign: P=0.92/R=0.92; malignant: P=0.91/R=0.91) and produces the fewest unnecessary referrals at that threshold. The two models have complimentary clinical roles: ViT-Small/16 is better when the probability of malignancy scores are needed to rank or stratify patients whereas EfficientNetV2-S provides the best balance in benign versus malignant classification when a fixed binary decision threshold is needed.

7. CONCLUSION

This study conducted a rigorous controlled comparison of eight binary skin lesion classifiers distinguishing benign from malignant dermoscopic lesions across Custom CNN (v1), ResNet18 (v2 and v3), DenseNet121, MobileNetV3-Large, ViT-Small/16, ConvNeXt-Tiny, and EfficientNetV2-S, all evaluated under identical dataset splits, preprocessing, and protocols on a 3,298 image dermoscopic dataset. Unlike most published studies that evaluate a single architecture under incomparable conditions, this work employs a unified experimental framework across eight models and prioritises ROC-AUC and PR-AUC as the primary evaluation criteria metrics that better reflect clinical screening utility than accuracy alone.EfficientNetV2-S (ImageNet-21K) delivers the strongest overall benign-malignant discrimination at a fixed decision threshold with 91.53% accuracy, macro-F1 of 0.907, malignant F1 of 0.910, ROC-AUC of 0.9639, and only 28 malignant lesions misclassified as benign from 300 malignant test cases a 9% false-negative rate. ViT-Small/16 achieves the highest ROC-AUC (0.9723) and highest malignant recall (0.91), making it the preferred model for sensitivity maximizing clinical screening where missing a malignant lesion carries the greatest patient risk. MobileNetV3-Large misclassifies 61 malignant lesions as benign a 20% false-negative rate confirming it is clinically unsafe for standalone malignant lesion screening. The two models serve complementary clinical roles: ViT-Small/16 for probability-ranked screening and risk stratification, and EfficientNetV2-S for fixed-threshold binary triage where minimising over-referral is equally important.

References

1. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
2. H. A. Haenssle et al., "Man against machine: diagnostic performance of a deep learning CNN for dermoscopic melanoma recognition," Ann. Oncol., vol. 29, no. 8, pp. 1836–1842, 2018.
3. M. A. Marchetti et al., "Results of the 2018 ISIC Challenge on AI for Skin Cancer Detection," JAMA Dermatol., vol. 155, no. 6, pp. 735–737, 2019.
4. P. Nayak et al., "Enhancing early skin cancer detection through AI-based image classification," Proc. ICOCT 2025, pp. 1–6.
5. O. Akinrinade and C. Du, "Skin cancer detection using deep machine learning techniques," Intelligence-Based Medicine, vol. 11, p. 100191, 2024.
6. P. S. Prakash Kumar et al., "Design and development of a ViT model for predicting skin cancer," Proc. ICICCS 2025, pp. 1493–1498.
7. K. B. Baabu and D. M. Raja S, "Skin cancer detection using Vision Transformer with GNN and explainable AI," Proc. AMATHE 2025, pp. 1–6.
8. A. El Mahdi et al., "Fusion of Vision Transformer and VGG features for enhanced skin lesion classification," Proc. ICCSC 2025, pp. 1–6.
9. A. Ali, H. Shahbaz, and R. Damasevicius, "xCViT: Improved Vision Transformer with CNN and Xception for skin disease recognition," Comput. Mater. Contin., vol. 79, no. 3, 2025.
10. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," Proc. ICLR 2021.
11. Z. Liu et al., "A ConvNet for the 2020s," Proc. CVPR 2022, pp. 11976–11986.
12. M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," Proc. ICML 2021, pp. 10096–10106.
13. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images," Sci. Data, vol. 5, p. 180161, 2018.
14. H. Naseri and A. A. Safaei, "Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review," BMC Cancer, vol. 25, p. 75, Jan. 2025.
15. C. Magalhães et al., "Deep Learning for Melanoma Detection: Comparing DenseNet121, ResNet50V2, NASNetMobile, and MobileNetV2 for binary classification," Diagnostics, vol. 15, no. 1, Jan. 2025.
16. Bello et al., "Skin Cancer Classification Using Fine-Tuned Transfer Learning of DenseNet-121," Applied Sciences, vol. 14, no. 17, p. 7707, 2024.
17. M. K. Uddin, "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.
18. I. Khoulqi et al., "Boosting Skin Cancer Classification: A Multi-Scale Attention and Ensemble Approach with Vision Transformers," Diagnostics, 2025.
19. Chiu et al., "Deep Ensemble Learning for Multiclass Skin Lesion Classification integrating Swin Transformer, ViT, and EfficientNetB4," Bioengineering, vol. 12, no. 9, p. 934, 2025.
20. S. S. Noronha et al., "Skin cancer detection using dermoscopic images with convolutional neural network (FCDS-CNN with EfficientNetV2-M)," Sci. Rep., vol. 15, Mar. 2025.
21. E. Baykal Kablan and S. Ayas, "Skin lesion classification from dermoscopy images using ensemble learning of ConvNeXt models," Signal Image Video Process., vol. 18, pp. 6353–6361, 2024.
22. Fernández et al., "Symmetry in Explainable AI: A Morphometric Deep Learning Analysis for Skin Lesion Classification," Symmetry, vol. 17, no. 8, p. 1264, 2025.