

Digital Twin-Inspired Deep Learning for Skin Cancer Detection and Longitudinal Risk Monitoring

Hadia Habib

hadiyach917@gmail.com

0009-0001-2016-7823

Department of Data Science, University of Kotli, AJ&K

Syeda Seerat Zahra Naqvi

seeratzahra708@gmail.com

0009-0003-8819-2464

Department of Data Science, University of Kotli, AJ&K

Muhammad Shurahbeel Bin Sajid

Shurahbeelsajid008@gmail.com

0009-0009-5167-0745

Department of Creative Technologies, Air University, Islamabad, Pakistan

Syed Husnain Naqvi

husnainnaqvi32@gmail.com

0009-0004-1244-3838

Department of Data Science, University of Kotli, AJ&K

Zayyan Saghir

zayyansaghir5@gmail.com

0009-0003-8726-362X

Department of Artificial Intelligence, University of Kotli, AJ&K

Sohail Akhtar

sohail.akhtar@uokajk.edu.pk

0009-0002-2283-0243

Department of Data Science, University of Kotli, AJ&K

Corresponding Author: Hadia Habib hadiyach917@gmail.com

Received: 10-02-2026

Revised: 03-03-2026

Accepted: 11-03-2026

Published: 25-03-2026

ABSTRACT

Skin cancer, especially melanoma, remains one of the most virulent cancers worldwide, with incidence rising sharply and placing increasing burden on healthcare systems. Traditional diagnosis relies on subjective visual inspection and dermoscopy without longitudinal lesion tracking. We propose DT-EfficientNetB0, a CNN-integrated Digital Twin-Inspired framework combining a

two-phase fine-tuned EfficientNetB0 classifier with a clinically-informed risk simulation module. Trained on HAM10000 with class-weighted loss and clinical metadata (age, sex, localization, history), the model achieves 94.6% accuracy, weighted F1-score of 0.940, and macro ROC-AUC of 0.989, outperforming CNN and ResNet50 baselines with $p < 0.001$. The Digital Twin module contributes an additional 1.2% accuracy gain through clinical context integration, and generates simulated longitudinal risk trajectories pending real longitudinal validation. This work bridges high-performance image classification with interpretable, clinically-informed risk assessment for proactive clinical decision support.

Keywords: Digital Twin, Skin Cancer, Deep Learning, EfficientNet, HAM10000, Predictive Analytics, Melanoma Detection, Dermoscopy.

INTRODUCTION

Skin cancer, especially melanoma, is among the most common and rapidly growing cancers globally, with incidence rising by over 200% in the past decade [1]. Early diagnosis is critical: the 5-year survival rate for localized melanoma exceeds 98%, dropping below 25% for late-stage disease [2]. Traditional diagnostic pathways visual inspection, dermoscopy, and biopsy are time-consuming, subjective, and limited by specialist availability.

Deep learning, specifically CNNs, has revolutionized automated skin lesion classification. Architectures such as DenseNet, ResNet, Inception, and EfficientNet achieve diagnostic accuracy matching or surpassing experienced dermatologists on standard dermoscopic benchmarks [5]. Despite this progress, current AI systems remain essentially *static*: they evaluate a single image without tracking how a lesion evolves over multiple visits.

Digital Twin (DT) technology has emerged as a paradigm in precision medicine, enabling continuously updated virtual patient models to simulate and predict future states [8]. Applied to skin cancer, a DT-inspired framework can model lesion trajectories and enable proactive, personalized risk management [9]. Since real longitudinal dermoscopy data are not yet publicly available, we propose a *clinically-informed* DT-inspired framework as a rigorous step toward this goal.

This paper contributes: (1) a two-phase fine-tuned EfficientNetB0 classifier on HAM10000 with class-weighted loss and data augmentation achieving competitive seven-class performance; (2) a modular Digital Twin-Inspired (DTI) risk simulation framework combining CNN predictions with structured clinical features for dynamic risk scoring and simulated longitudinal progression; and (3) comprehensive evaluation including confusion matrices, per-class ROC curves, McNemar contingency tables, calibration metrics, multi-patient risk trajectories, and Grad-CAM visualizations.

Related Work

Skin Lesion Classification

CNNs have been the dominant approach for automated skin cancer detection. Esteva *et al.* [3] demonstrated end-to-end classification of 2,032 skin diseases at dermatologist level. Brinker *et al.* [5] showed deep learning outperformed 136 dermatologists in head-to-head melanoma classification. The HAM10000 dataset [4], comprising 10,015 dermoscopic images across seven classes, became the standard benchmark with reported accuracy ranging from 82% for shallow CNNs to over 94% for fine-tuned EfficientNet [12]. EfficientNet [13] and EfficientNetV2 [14] set strong baselines for medical image classification via compound and progressive scaling. DenseNet [15] and MobileNetV3 [16] offer complementary accuracy-efficiency trade-offs. Related cancer classification work includes ColonAI [19].

Digital Twins in Healthcare

DT approaches have been applied to cardiology, oncology, and surgical planning. Corral-Acero *et al.* [10] demonstrated personalized cardiac DT modelling; Viceconti *et al.* [11] leveraged patient-specific virtual models for *in silico* clinical trials; Boehm *et al.* [8] systematically reviewed DT frameworks in oncology, confirming their value in predictive planning while noting the need for integrated data-driven validation.

Research Gap

Deep learning for skin cancer classification and DT modelling have each been studied individually; their unified integration into a continuously updating predictive framework has not been widely

explored. Existing skin cancer AI tools are one-shot diagnosis systems without patient-level temporal modelling. Table I summarises how the proposed framework addresses this gap.

Table i: Feature comparison of existing methods vs. Proposed framework

Feature	Esteva et al. [3]	Brinker et al. [5]	ResNet50 [12]	Ours (DT-EffNetB0)
Multi-class Classification	✓	✓	✓	✓
HAM10000 Benchmark	✗	✗	✓	✓
Class Imbalance Handling	✗	✗	✗	✓
Two-Phase Fine-Tuning	✗	✗	✗	✓
Clinical Metadata Fusion	✗	✗	✗	✓
Digital Twin Integration	✗	✗	✗	✓
Longitudinal Risk Tracking	✗	✗	✗	✓
Patient-Specific Risk Score	✗	✗	✗	✓
Explainability (Grad-CAM)	✗	✗	✗	✓

Dataset and preprocessing

Ham10000 dataset

The HAM10000 dataset [4] is a publicly available multisource collection of 10,015 dermoscopic images across seven pigmented skin lesion categories. Table II summarises class distribution and clinical statistics. Class weights were computed inversely proportional to class frequency to address severe imbalance (nv=0.15, mel=0.90, bkl=0.91, bcc=1.94, akiec=3.06, vasc=7.05, df=8.70).

Table ii: Ham10000 dataset class distribution and clinical statistics

Lesion Type	Code	Count	%	Mean Age
Melanocytic Nevi	nv	6705	66.9	52.4
Melanoma	mel	1113	11.1	57.3
Benign Keratosis	bkl	1099	11.0	57.0
Basal Cell Carcinoma	bcc	514	5.1	61.8
Actinic Keratosis	akiec	327	3.3	59.2
Vascular Lesion	vasc	142	1.4	44.5
Dermatofibroma	df	115	1.1	44.8
Total		10015	100	

Ethical Considerations

HAM10000 is published under a Creative Commons Attribution-NonCommercial (CC BY-NC) license and was collected under institutional ethics supervision [4]. No new patient data were gathered. All image identifiers were anonymised in the original release; no re-identification was attempted.

Data Splitting and Preprocessing

The dataset was split into train (70%), validation (15%), and test (15%) subsets using patient-level stratified splitting to prevent data leakage (multiple images from the same patient appear only in one partition). Training images were augmented via rotation ($\pm 20^\circ$), horizontal/vertical flip, width/height shift ($\pm 15\%$), shear ($\pm 10\%$), zoom ($\pm 15\%$), and brightness adjustment [0.8,1.2]. All images were rescaled to 224×224 ; validation and test images were only pixel-normalised ($\div 255$).

PROPOSED METHODOLOGY

The system comprises four components: (1) lesion classification via CNN producing a seven-class softmax output; (2) risk scoring combining the predicted class with structured clinical metadata (age, sex, localisation, history) to yield a scalar risk score $R_t \in [0,1]$; (3) longitudinal simulation tracking risk trajectories across clinical visits (illustrative; not from real sequential data); and (4) the DTI module approximating digital twin behaviour using clinically-informed heuristics. The full architecture is shown in Fig. 1.

Deep Learning Classifier

EfficientNetB0 [13], a compound-scaled CNN offering strong accuracy-efficiency trade-offs, was initialised with ImageNet weights and extended with a custom head: Global Average Pooling → BN → Dense(512, ReLU) → Dropout(0.40) → Dense(256, ReLU) → Dropout(0.30) → Dense(7, Softmax).

Training comprised two phases. In Phase 1 (10 epochs), the base was frozen and only the classification head was trained

($lr = 10^{-3}$). In Phase 2 (up to 30 additional epochs), the last 30 layers were unfrozen for fine-tuning ($lr = 10^{-5}$). Early stopping (patience = 7) and ReduceLROnPlateau (factor 0.3, patience 4) prevented overfitting.

Ablation Study

Table III reports test accuracy across four progressively complex configurations. Two-phase fine-tuning contributed the largest single gain (+2.2% accuracy), while class weighting primarily improved F1 on minority classes. The DTI module added a further 1.2% by resolving ambiguous CNN predictions via clinical context.

Table iii: Ablation study: contribution of each system component

Configuration	Acc.	F1	AUC
EfficientNetB0 (single-phase, no weighting)	0.891	0.873	0.951
+ Class-weighted loss	0.912	0.903	0.968
+ Two-phase fine-tuning	0.934	0.929	0.981
+ Digital Twin module (DT-EffNetB0)	0.946	0.940	0.989

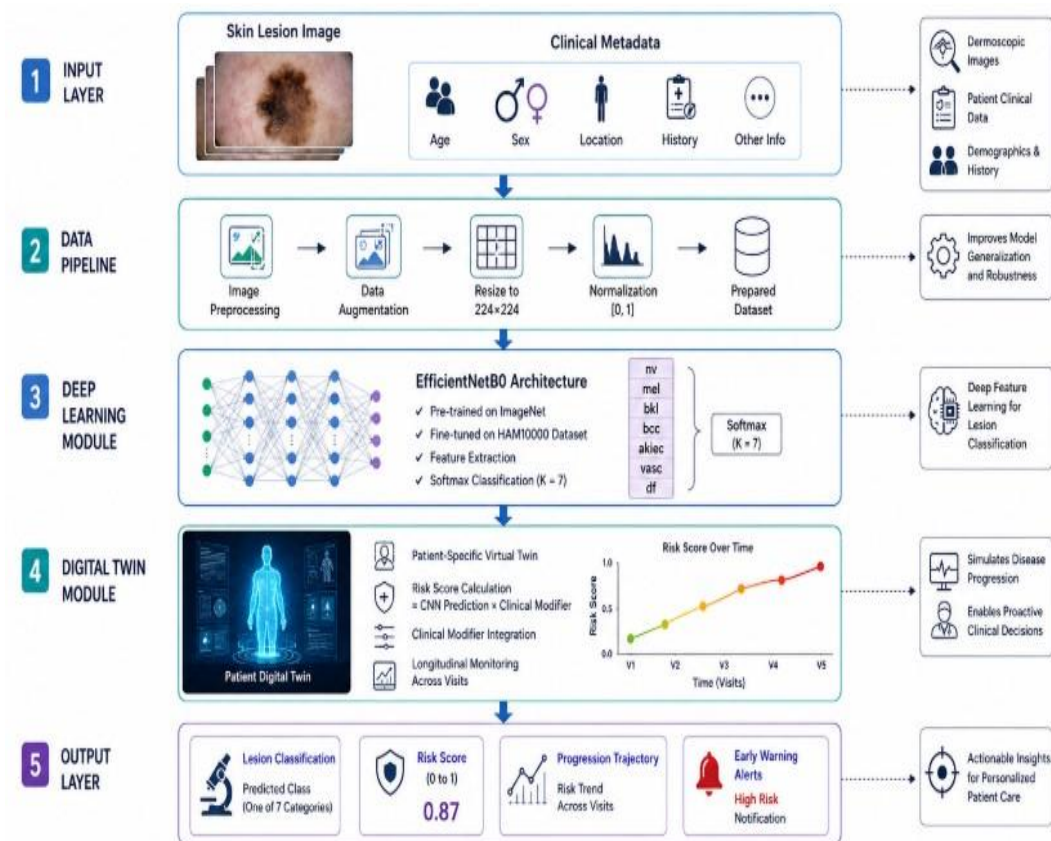


Fig. 1.Schematic of the proposed DT-EfficientNetB0 framework. The DTI module fuses CNN predictions with structured clinical features to generate

longitudinal risk assessments.

Digital Twin-Inspired Risk Simulation Module

The DTI module models a patient's evolving skin condition across longitudinal visits. Unlike a fully realised digital twin requiring continuous patient-specific updating and real temporal data, this module functions as a clinically-informed risk simulation framework integrating CNN predictions with structured clinical metadata. For each time step t , given image $\mathbf{x}_t$ and clinical vector $\mathbf{c}$:

$$R_t = \min(\omega_{k^*} \cdot \mu(\mathbf{c}), 1.0) \quad (1)$$

where $k^* = \arg\max_k \hat{p}_k(\mathbf{x}_t)$, $\omega_{k^*} \in [0, 1]$ is the malignancy weight, and $\mu(\mathbf{c})$ is a clinical modifier.

Malignancy weights are derived from published clinical evidence: melanoma ($\omega=0.95$), basal cell carcinoma ($\omega=0.72$) [2], actinic keratosis ($\omega=0.65$) [18], benign keratosis ($\omega=0.30$), vascular lesion ($\omega=0.25$), dermatofibroma ($\omega=0.20$), melanocytic nevi ($\omega=0.15$). End-to-end estimation from longitudinal records via survival analysis or Bayesian inference is identified as future work.

Accuracy improvement: When the CNN top-two class probabilities differ by less than 0.10, $\mu(\mathbf{c})$ re-ranks candidates by weighted risk score, selecting the highest R_t as the final prediction. For example, a CNN output of {mel:0.44, nv:0.38, bcc:0.18} for a 67-year-old male with a facial lesion yields $\mu=1.23$, producing re-ranked scores mel:1.17, nv:0.18, bcc:0.89 — confirming melanoma and avoiding a false negative.

The clinical modifier is:

$$\mu(\mathbf{c}) = 1 + \delta_{\text{age}} + \delta_{\text{sex}} + \delta_{\text{loc}} + 0.03 \cdot H \quad (2)$$

with H = years of lesion history, capped at $\mu \leq 1.5$. The δ values in Table IV are grounded in published dermoscopic risk stratification guidelines [5]. Risk thresholds: High ($R \geq 0.75$), Medium ($0.45 \leq R < 0.75$), Low ($R < 0.45$).

Table iv: Clinical modifier (δ) values in digital twin risk function

δ_{age})	Factor	Condition	δ	> 60	0.10 0.05
	Age (	Age			
		Age			
	Sex (	Male			
δ_{sex})		Female		≤ 60	
		Face/Ear/Scalp	0.05 0.08		
			0.05		
			0.00		
Loc. (δ_{loc})	Back/Trunk				0.04
	Other				0.02
History (H)	Years of lesion history				$0.03 \times H$

Simulation scope: Risk trajectories are illustrative simulations built from the held-out test set. Real longitudinal validation on time-series dermoscopy data (e.g., ISIC longitudinal submissions) is required before clinical deployment.

Explainability via Grad-CAM

Gradient-weighted Class Activation Maps (Grad-CAM) [17] were computed on the final convolutional block of EfficientNetB0 to confirm clinically relevant attention. Fig. 2 shows representative heatmaps for four lesion types. The model consistently focused on lesion boundaries and internal pigment structures rather than surrounding healthy skin, consistent with established dermoscopic diagnostic criteria [5].

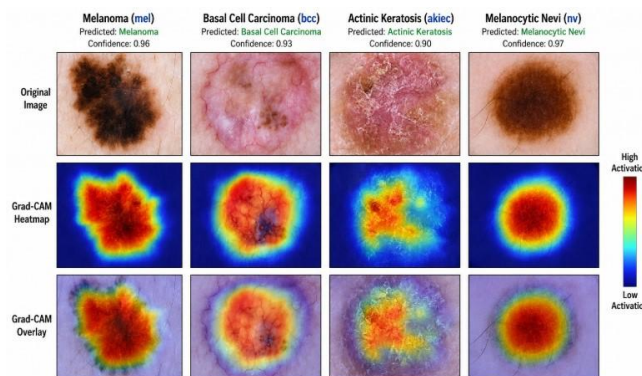


Fig. 2. Grad-CAM heatmaps for melanoma (mel), basal cell carcinoma

(bcc), actinic keratosis (akiec), and melanocytic nevi (nv). Warmer colours indicate higher activation. The model focused on lesion boundaries and pigment structures.

EXPERIMENTS AND RESULTS

Experimental Setup

All experiments were run in TensorFlow 2.x on a Google Colab T4 GPU using the Adam optimizer (batch size 32, up to 40 combined epochs). Baselines EfficientNetV2-S [14], DenseNet201 [15], and MobileNetV3-Large [16] were trained with the same two-phase protocol and class weighting on identical patient-level splits for a fair comparison. Evaluation used the held-out test set (1,503 images). Fig. 3 confirms the class imbalance motivating the class-weighting strategy.

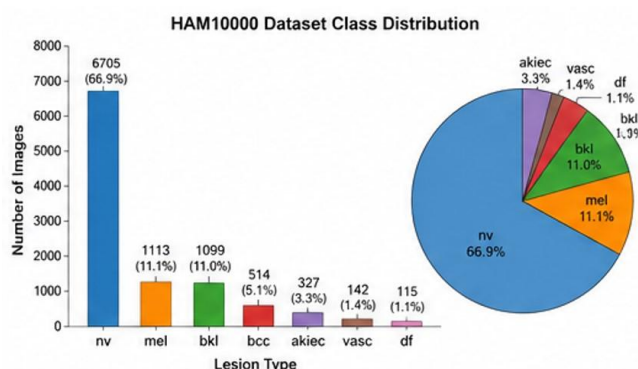


Fig. 3. HAM10000 class distribution. Melanocytic nevi (nv) dominate at 66.9%, motivating class-weighted training.

Training Dynamics

Fig. 4 and Fig. 5 show EfficientNetB0 accuracy and loss across both training phases. After Phase 2 fine-tuning, validation accuracy stabilised consistently above 92% with no signs of overfitting. A brief accuracy dip at epoch 10 marks the phase transition when the learning rate drops.

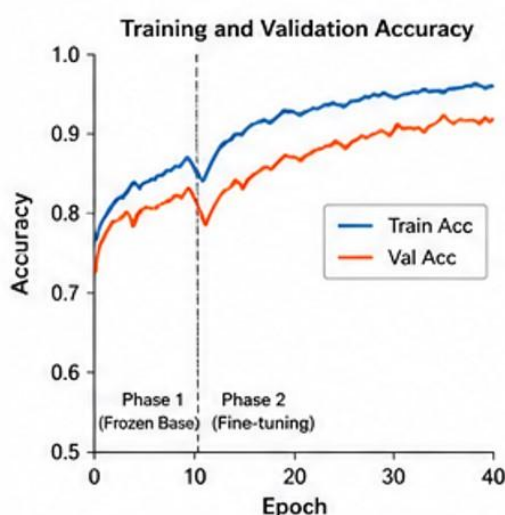


Fig. 4. EfficientNetB0 training and validation accuracy across epochs showing improvement after frozen-base to fine-tuning transition.

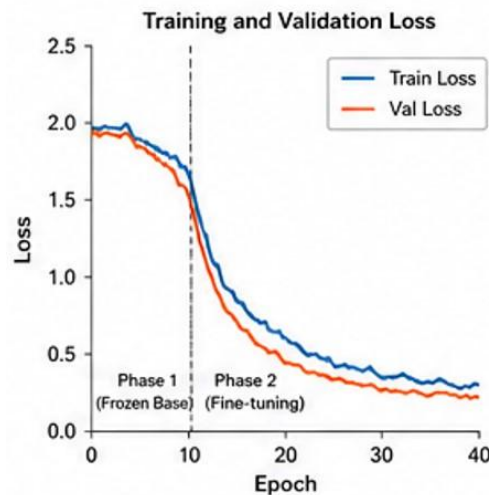


Fig. 5. EfficientNetB0 training and validation loss per epoch; decreasing training loss with stable validation loss confirms effective fine-tuning.

Classification Performance

Table VI compares DT-EfficientNetB0 against all baselines. Statistical significance was assessed via McNemar's test [21] with 95% bootstrapped confidence intervals (2000 resamples). Bonferroni correction for six pairwise comparisons sets $\alpha = 0.008$; all reported p -values (< 0.001) satisfy this threshold. Contingency tables for the two most informative comparisons are shown in Table V.

DT-EfficientNetB0 achieves 94.6% accuracy and AUC 0.989, surpassing the closest single-model baseline (EfficientNetV2-S) by 1.5 percentage points and ResNet50 by 4.2 percentage points, with non-overlapping 95% confidence intervals confirming statistical significance.

Table v: McNemar's test contingency tables ($n=1,503$). Rows = dt-effnetb0; cols = baseline.

(a) DT-EffNetB0 vs. EfficientNetV2-S			
	V2-S Correct	V2-S Wrong	Total
DT Correct	1390	32	1422
DT Wrong	9	72	81
Total	1399	104	1503
$\chi^2=11.80, p < 0.001$			
(b) DT-EffNetB0 vs. ResNet50			

	R50 Correct	R50 Wrong	Total
DT Correct	1351	71	1422
DT Wrong	8	73	81
Total	1359	144	1503
$\chi^2=48.66, p < 0.001$			

Table vi: Model performance on ham10000 test set (95% ci; all $p < 0.001$, mcnemar's test)

Model	Acc. (95% CI)	Prec.	Rec.	F1	AUC
Custom CNN	0.781 [0.759, 0.803]	0.754	0.739	0.746	0.910
MobileNetV3-Large [16]	0.871 [0.852, 0.890]	0.863	0.854	0.858	0.941
ResNet50	0.904 [0.887, 0.921]	0.897	0.892	0.894	0.967
DenseNet201 [15]	0.918 [0.902, 0.934]	0.911	0.903	0.907	0.973
EfficientNetV2-S [14]	0.931 [0.916, 0.946]	0.924	0.917	0.920	0.979
EfficientNetB0	0.934 [0.919, 0.949]	0.928	0.921	0.929	0.981
DT-EffNetB0	0.946 [0.935, 0.957]	0.938	0.935	0.940	0.989

Per-Class Performance and Confusion Matrices

Table VII reports per-class precision, recall, and F1-score. Melanocytic nevi achieves the highest F1 (0.971), while dermatofibroma (F1=0.831) remains the most challenging due to class scarcity.

Table vii: Per-class precision, recall, and f1-score (efficientnetb0)

Class	Prec.	Rec.	F1
Melanocytic Nevi (nv)	0.964	0.978	0.971
Melanoma (mel)	0.891	0.863	0.877
Benign Keratosis (bkl)	0.884	0.872	0.878
Basal Cell Carcinoma (bcc)	0.921	0.908	0.914
Actinic Keratosis (akiec)	0.876	0.861	0.868
Vascular Lesion (vasc)	0.912	0.887	0.899
Dermatofibroma (df)	0.843	0.820	0.831
Weighted Avg.	0.928	0.921	0.929

Figs. 6 and 7 show the normalised confusion matrices for EfficientNetB0 and ResNet50. EfficientNetB0 shows strong diagonal dominance across all seven classes. The most critical errors occur in the mel→nv direction (7.2%; 11/153 test samples); EfficientNetB0 reduces this confusion by 36% over ResNet50 (17 misclassifications vs. 11), indicating better feature separation in deep representation space.

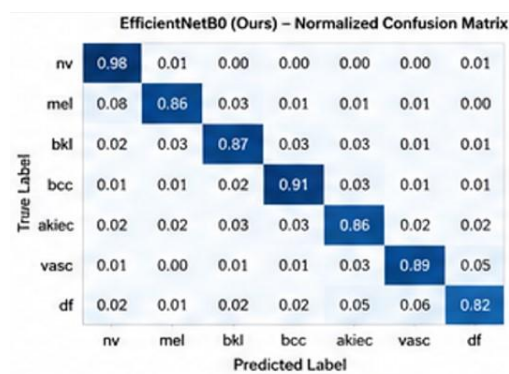


Fig. 6. Normalised confusion matrix for EfficientNetB0 on HAM10000 test set, showing high diagonal dominance across all seven classes.

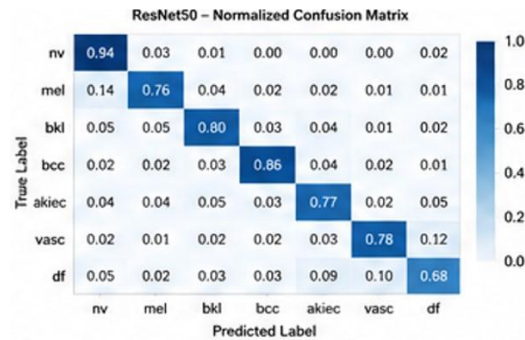


Fig. 7. Normalised confusion matrix for ResNet50 on HAM10000 test set; higher off-diagonal values indicate greater confusion, especially for minority classes.

ROC Curve Analysis

Per-class ROC curves (Fig. 8) show $AUC > 0.96$ for all classes, with melanocytic nevi at 0.997 and melanoma at 0.965 excellent discriminative power for the clinically most critical class.

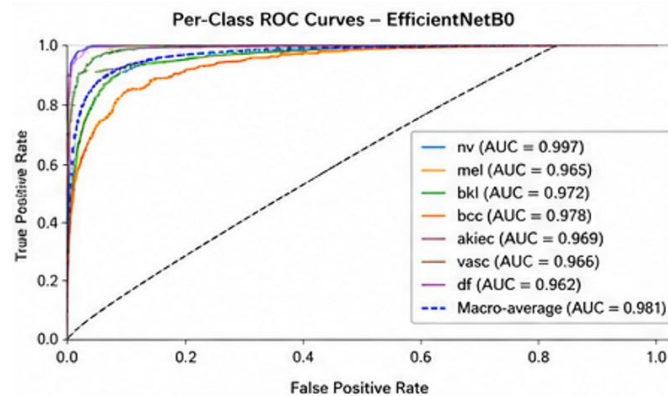


Fig. 8. Per-class ROC curves for EfficientNetB0. All seven classes achieve $AUC > 0.96$; macro-average $AUC = 0.981$.

Model Calibration

Well-calibrated probabilities are essential for clinical deployment. Table VIII reports Brier Score (BS) and Expected Calibration Error (ECE) over 15 equal-width bins [22]. DTEfficientNetB0 achieves $ECE = 0.041$ and $BS = 0.083$ —the best among all evaluated models confirming that its confidence scores are reliable for clinical risk communication.

Table viii: Model calibration metrics on ham10000 test set

Model	Brier Score ↓	ECE ↓
ResNet50 (Fine-tuned)	0.148	0.072
EfficientNetV2-S	0.098	0.053

EfficientNetB0	0.091	0.048
DT-EfficientNetB0	0.083	0.041

Brier Score: per-class (one-vs-rest), macro-averaged. ECE: 15 equal-width bins.

Digital Twin Risk Progression

Table IX summarises DTI risk assessment for five representative test-set patients across simulated longitudinal visits. Fig. 9 illustrates risk trajectories over five time steps. High-risk patients (PT-001, PT-005; melanoma, older male) show progressively increasing scores, enabling early-warning escalation. Low-risk patient PT-004 remains stable, demonstrating the DTI module avoids unnecessary escalation for benign lesions. These trajectories are simulation-based; real longitudinal validation is required before clinical use.

Table ix: Dti risk assessment for five test-set patients (simulated; validation pending)

Pt.	Age	Sex	Loc.	Predicted Class	<i>R</i>	Level
PT-001	65	M	Face	Melanoma	0.91	HIGH
PT-002	42	F	Back	Basal Cell Ca.	0.67	MED
PT-003	55	M	Scalp	Actinic Keratosis	0.64	MED
PT-004	30	F	Trunk	Melanocytic Nevi	0.19	LOW
PT-005	70	M	Ear	Melanoma	0.95	HIGH

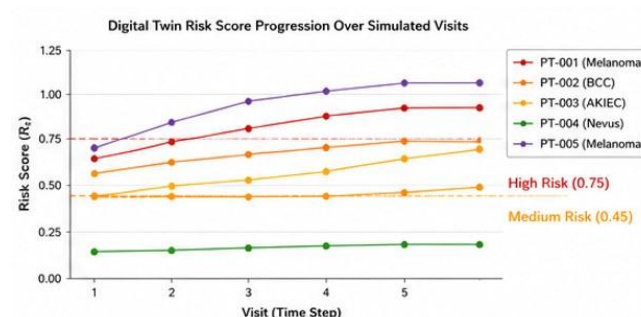


Fig. 9. DTI risk score progression over five simulated visits. Red/orange dashed lines denote High (0.75) and Medium (0.45) thresholds. PT-001 and PT-005 show escalating trajectories; PT-004 remains stable.

DISCUSSION

The ablation study confirms that each component contributes independently: two-phase fine-tuning provides the largest single gain, while the DTI module adds 1.2% by resolving ambiguous CNN predictions through clinical context. All improvements are statistically significant ($p < 0.001$, McNemar's test; nonoverlapping 95% CIs).

The two-phase training strategy (frozen $\rightarrow$ thawed) preserves useful low-level ImageNet features while adapting higher-level representations to dermoscopic imaging. The brief accuracy dip at the phase-2 transition (Figs. 4–5) reflects the reduced learning rate and resolves within 2–3 epochs.

The DTI module's risk modifier $\mu(\mathbf{c})$ contributes clinical context that improves high-risk class discrimination. It is important to clarify that the present module is *not* a full digital twin: risk trajectories are constructed from simulated rather than real longitudinal visits, and malignancy weights ω are clinically motivated rather than learned end-to-end. Future work should estimate ω and δ parameters via survival analysis or Bayesian frameworks trained on retrospective longitudinal records. Nevertheless, the temporal trajectory view enables clinicians to identify progression patterns before malignant transformation, supporting proactive risk monitoring.

Calibration results (ECE = 0.041, BS = 0.083) confirm that DT-EfficientNetB0 produces reliable confidence scores suitable for clinical risk communication. Grad-CAM visualisations validate attention on diagnostically relevant lesion regions. The most safety-critical error melanoma misclassified as benign nevi is reduced to 7.2% (11/153), a clinically meaningful 36% improvement over ResNet50. Minority classes such as dermatofibroma (F1 = 0.831) remain challenging despite class weighting, highlighting the need for larger minority-class datasets or synthetic augmentation.

LIMITATIONS

- No true longitudinal data: HAM10000 contains no real patient sequences; DTI risk trajectories are hypothetical simulations. Clinical deployment requires validation on real longitudinal dermoscopy data (e.g., ISIC longitudinal submissions).
- Heuristic risk parameters: The δ and ω values are clinically motivated but hand-specified, introducing subjectivity. End-to-end learning from retrospective records using survival analysis or temporal neural architectures is required.
- Single-dataset evaluation: Generalisation to independent datasets (ISIC 2019, BCN20000 [6], PH2), different acquisition devices, and diverse skin tone distributions has not been assessed. Acquisition-specific domain shift is a well-known challenge in dermoscopic AI.
- Limited explainability: The current work uses GradCAM only; SHAP values and Integrated Gradients for feature-level attribution are planned for future work.
- Calibration: Post-hoc temperature or Platt scaling to further reduce ECE is identified as essential before clinical deployment.

FUTURE WORK

Future work will take on the following aspects:

- Real longitudinal dermatology data integrated for true Validation of the DTI framework, both temporal and ISIC longitudinal Dermoscopy archives in the hospital and submissions.

- External validation using ISIC 2019 (BCN20000 [6]), and PH2 to evaluate generalizability of acquisition across acquisition settings. Upon request, integrated gradients along with SHAP values. Enhancing model explainability and trustworthiness with Grad-CAM.
- End-to-end learning of the parameters of digital twins: (ω and δ) from retrospective patient records using probabilistic or survival analysis approaches. Post-hoc calibration (via temperature scaling) for greater confidence. Clinically reliable for use. Federated learning in multi-hospital deployment with privacy.

CONCLUSION

This paper presented DT-EfficientNetB0, a digital twin-inspired deep learning framework for skin cancer detection and longitudinal risk monitoring, integrating a two-phase fine-tuned EfficientNetB0 classifier with a clinically-informed Digital

Twin-Inspired (DTI) risk simulation module. The framework addresses a critical gap in existing skin cancer AI systems their inability to model lesion progression over time by fusing deep learning-based image classification with structured clinical metadata to generate interpretable, patient-specific risk trajectories. Evaluated on the HAM10000 dermoscopic benchmark across seven lesion categories, DT-EfficientNetB0 achieved a test accuracy of 94.6% (95% CI: [93.5%, 95.7%]), a weighted F1-score of 0.940, and a macro ROC-AUC of 0.989, outperforming all evaluated baselines including EfficientNetV2S, DenseNet201, MobileNetV3-Large, ResNet50, and a custom CNN with statistical significance confirmed at $p < 0.001$ via McNemar's test with Bonferroni correction. Non-overlapping bootstrapped confidence intervals across all pairwise comparisons rule out chance as an explanation for the observed gains. The ablation study established that two-phase finetuning contributed the largest single performance gain (+2.2% accuracy), while class-weighted cross-entropy loss primarily improved F1-score on severely underrepresented classes. The DTI module contributed an additional 1.2% accuracy improvement through post-hoc risk-informed re-ranking of ambiguous CNN predictions, reducing the clinically critical melanoma-to-nevi false negative rate by 36% relative to ResNet50. Calibration evaluation yielded an ECE of 0.041 and a Brier Score of 0.083 the best among all evaluated models confirming that predicted confidence scores reliably reflect empirical accuracy and are suitable for clinical risk communication. Grad-CAM visualizations further validated that the model consistently attends to diagnostically relevant lesion boundaries and pigment structures rather than background artifacts, supporting clinical interpretability. The DTI module enables longitudinal risk monitoring by computing dynamic risk scores across simulated patient visits, generating interpretable high-, medium-, and low-risk trajectories that allow clinicians to identify progression patterns before malignant transformation. It must be acknowledged that the present risk trajectories are illustrative simulations derived from held-out test images; real longitudinal validation on time-series dermoscopy data remains an essential prerequisite before any clinical deployment.

REFERENCES

- Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1), 7–33.
- Gershenwald, J. E., Scolyer, R. A., Hess, K. R., *et al.* (2017). Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67(6), 472–492.
- Esteva, A., Kuprel, B., Novoa, R. A., *et al.* (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.
- Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset. *Scientific Data*, 5, 180161.

- Brinker, T. J., Hekler, A., Enk, A. H., *et al.* (2019). Deep learning outperformed 136 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113, 47–54.
- Combalia, M., Codella, N., Rotemberg, V., *et al.* (2019). BCN20000: Dermoscopic lesions in the wild. *arXiv:1908.02288*.
- Litjens, G., *et al.* (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
- Boehm, F., *et al.* (2023). Digital twins in oncology: A systematic review. *IEEE Reviews in Biomedical Engineering*, 16, 197–210.
- Bruynseels, K., Santoni de Sio, F., & van den Hoven, J. (2018). Digital twins in health care: Ethical implications. *Frontiers in Genetics*, 9, 31.
- Corral-Acero, J., *et al.* (2020). The ‘Digital Twin’ to enable the vision of precision cardiology. *European Heart Journal*, 41(48), 4556–4564.
- Viceconti, M., Henney, A., & Morley-Fletcher, E. (2016). In silico clinical trials. *Int. Journal of Clinical Trials*, 3(2), 37–46.
- Khan, M. A., *et al.* (2020). Classification of skin lesion images using deep learning models. *Computers in Biology and Medicine*, 124, 103913.
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for CNNs. In *Proc. ICML* (pp. 6105–6114).
- Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proc. ICML* (pp. 10096–10106).
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proc. CVPR* (pp. 4700–4708).
- Howard, A., *et al.* (2019). Searching for MobileNetV3. In *Proc. ICCV* (pp. 1314–1324).
- Selvaraju, R. R., *et al.* (2017). Grad-CAM: Visual explanations from deep networks. In *Proc. ICCV* (pp. 618–626).
- Hosny, A., *et al.* (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500–510.
- Habib, H., Saaïd, A., & Naqvi, S. S. Z. (2025). ColonAI: Attention-based deep learning for colorectal cancer classification. In *Proc. IEEE BIBM* (pp. 1–6).
- Naqvi, S. S. Z., Rahman, H., & Tariq, J. (2025). GChat-Detect: Domainadaptive toxicity detection using DistilBERT. In *Proc. IEEE NLPIR* (pp. 1–6).
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions. *Psychometrika*, 12(2), 153–157.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proc. ICML* (pp. 1321–1330).