International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

TriDermCancerNet: An Explainable Multi-Branch Hybrid Deep Learning Framework with Adaptive Feature Fusion for Robust Multi-Class Skin Cancer Classification

Riaz Ahmed
Associate Professor,
Department of Computer Applications,
Government Degree College, Rajouri, Jammu and Kashmir, India

Abstract - Background: Skin cancer is the most common type of cancer, accounting for one in three cancers diagnosed globally. Although melanoma represents just around 1% of skin cancer incidence, it accounts for the bulk of skin cancer deaths. Deep learning based dermoscopic image classification provides a scalable low-cost triage tool. Current single-branch CNN or Transformer models fail to simultaneously capture fine-grained texture (pigment networks, dots, streaks), global lesion morphology, and frequency-domain artefacts (hair, ruler marks, vignetting), limiting robustness across the seven-to-eight clinically relevant lesion categories. Methods: We present TriDermCancerNet, a multi-branch hybrid deep learning framework comprising three concurrent, architecturally different branches: (B1) a ConvNeXt-Tiny CNN branch to learn hierarchical texture and local lesion-boundary features, (B2) a Swin-Transformer-Tiny branch to learn global structural and shape context features with shifted-window self-attention, and (B3) a Discrete Wavelet Transform (DWT)-CNN branch in the frequency domain to suppress hair/artefact noise and enhance pigmentation-frequency signatures. An Adaptive Cross-Branch Attention Fusion (ACAF) module is proposed to dynamically reweight and fuse the three branch embeddings for each sample, through a learnt gating network conditioned on the agreement between the branches. Dual explain ability modules based on Grad-CAM++ and SHAP provide lesion-localized and feature-attribution based explanations, respectively. Experimental Setup: We evaluate three public ceroscopy benchmarks, HAM10000 (10,015 images, 7 classes), ISIC 2019 (25,331 images, 8 classes) and PH^2 (200 images, 3 classes) and a held-out clinical test set (1,840 images) from a tertiary dermatology centre in a five-fold stratified cross-validation. Results: TriDermCancerNet yields 96.84% accuracy, 0.987 macro-AUC, 95.62% macro-F1, and 94.31% balanced multi-class accuracy on HAM10000 outperforming 12 single-branch and dual-branch baselines. Melanoma sensitivity is 97.3% (vs. 92.1% best baseline) at 95.8% specificity. The ACAF module outperforms the naive concatenation fusion by 2.14% in terms of accuracy, and the model generalizes to the external clinical test set with just 2.87% accuracy reduction. Conclusion: TriDermCancerNet offers strong, clinically explainable multi-class skin cancer classification with potential for inclusion into teledermatology triage pipelines via adaptive fusion of spatial texture, global structure, and frequency domain signal.

Keywords: Skin cancer classification; Melanoma detection; Multi-branch deep learning; Feature fusion; Vision Transformer; Discrete Wavelet Transform; Explainable AI; Dermoscopy; Grad-CAM++; SHAP

I. Introduction

Skin cancer is the most common cancer worldwide. The three main malignant types are melanoma, basal cell carcinoma (BCC) and squalors cell carcinoma (SCC). Several benign mimickers also exist, including melanocytic nevi, benign keratosis, dermato-fibroma and vascular lesions. Melanoma, although representing just 1% of skin cancers, is responsible for most skin-cancer mortality because of its high spreading potential, highlighting the clinical need for early and precise identification [1]. Automated techniques for triaging worrisome lesions for urgent assessment by a dermatologist provide a high-value clinical intervention, as the five-year survival rate for melanoma diagnosed at Stage I is greater than 99%, and drops to less than 30% for Stage IV cancer [2].

Polarized, magnified surface microscopy, or dermoscopy, visualizes sub-surface pigment structures not seen by the naked eye and is the clinical standard for lesion assessment. Expert visual diagnosis is guided by the ABCDE criteria (Asymmetry, Border irregularity, Color variation, Diameter, Evolution) and 7-point checklist, but inter-rater agreement between dermatologists is still only moderate (kappa ≈ 0.6–0.7) [3]. Computer-aided diagnosis (CAD) systems can provide consistent, reproducible second opinions, motivating their development. Since Esteva et al. [4] achieved dermatologist-level performance with Inception-v3 on a dataset of 129,450 images, deep learning has greatly improved dermoscopic CAD. However, most of the following designs rely on a single CNN or Transformer backbone, automatically emphasizing either local texture (CNNs, through limited receptive fields) or global structure (Transformers, through self-attention), but rarely both at the same time. Furthermore, dermoscopic pictures are often contaminated with confounding artefacts, such as terminal hair, air bubbles, ruler lines, ink annotations and vegetating from the dermoscope lens, which traditional spatial-domain models struggle to untangle from true lesion information.

We hypothesize that multi-branch architectures that explicitly model complementary signal types – local texture (CNN), global structure (Transformer) and frequency-domain artefact-robust features (Discrete Wavelet Transform) fused via learned adaptive fusion can outperform single-paradigm architectures while remaining explainable. This encourages the TriDermCancerNet. Our contributions are:

1. TriDermCancerNet architecture: a novel three-branch hybrid architecture combining ConvNeXt-Tiny (texture), Swin-Transformer-Tiny (global structure), and a DWT-CNN (frequency-domain artifact suppression) for dermoscopic image classification.

2. Adaptive Cross-Branch Attention Fusion (ACAF): a learnt sample-wise gating technique that dynamically re-weights the contributions of branches according to the inter-branch embedding agreement, improving over static concatenation and fixed-weight fusion.

3. Combination of Grad-CAM++ lesion localization maps and SHAP-based feature attribution in a dual explain ability pipeline tested against dermatologist ABCDE annotations.

4. Thorough evaluation on three public benchmarks and an independent clinical test set, compared to twelve baselines, thorough ablation study, and statistical significance testing.

II. Related Work

2.1 CNN-Based Skin Cancer Classification

Esteva et al. [4] pioneered dermatologist-level CNN classification using Inception-v3 on a 129,450-image proprietary dataset spanning 2,032 diseases. Tschandl et al. [5] released HAM10000, becoming the standard public benchmark; subsequent work applied DenseNet [6], EfficientNet [7], and ResNeXt [8] with progressive accuracy improvements. Attention-augmented CNNs [9] and multi-scale feature pyramid networks [10] further improved fine-grained lesion boundary discrimination. However, CNN receptive fields remain inherently local, limiting holistic lesion-shape reasoning critical to the Asymmetry and Border criteria of ABCDE.

2.2 Vision Transformers in Dermatology

Vision Transformer (ViT) [11] and its hierarchical variant Swin Transformer [12] introduced self-attention mechanisms capturing long-range dependencies. Xin et al. [13] applied ViT to HAM10000, achieving competitive accuracy but requiring substantially more training data than CNNs to converge. Swin-based hybrid models [14] addressed this via window-based local attention with hierarchical merging, improving data efficiency. Pure Transformer models, however, can under-emphasise fine local texture cues such as pigment network regularity, which dermatologists rely on heavily for melanoma discrimination.

2.3 Frequency-Domain and Artefact-Robust Methods

Dermoscopic images are frequently degraded by hair occlusion, present in an estimated 30–40% of clinical images [15]. Hair removal pre-processing (DullRazor, morphological closing) is a common but imperfect mitigation. Frequency-domain approaches—Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) features—have shown promise for artefact-robust texture characterisation in other medical imaging domains [16,17], but their integration into end-to-end skin cancer deep learning pipelines remains limited to a handful of preliminary studies [18].

2.4 Multi-Branch and Fusion Architectures

Multi-branch fusion has been explored for combining clinical metadata with image features [19], combining dermoscopic and clinical macroscopic images [20], and assembling multiple CNN backbones via late-fusion voting [21]. Most fusion strategies use static concatenation or fixed-weight averaging, which cannot adapt to per-sample reliability differences across branches. Adaptive or learned fusion gating, common in multimodal NLP-vision tasks [22], has not been systematically explored for intra-modal (same-image, different-representation) dermoscopic fusion.

2.5 Explainability in Dermoscopic AI

Grad-CAM [23] and Grad-CAM++ [24] are the dominant visual explanation tools in dermatology CAD, validated against expert lesion annotations in several studies [25]. SHAP [26] provides complementary feature-level attribution but is rarely combined with visual saliency in a unified pipeline. Regulatory guidance (FDA SaMD, EU MDR) increasingly requires multi-modal explainability for high-risk diagnostic AI, motivating our dual Grad-CAM++/SHAP approach.

2.6 Research Gap

No prior work combines CNN, Transformer, and frequency-domain branches with learned adaptive fusion specifically for dermoscopic multi-class classification, nor pairs this with dual visual/attribution explainability and rigorous external clinical validation. TriDermCancerNet addresses this combined gap.

III. Proposed Methodology: TriDermCancerNet

3.1 Architecture Overview

TriDermCancerNet processes each 384×384 dermoscopic input image through three parallel branches whose embeddings are merged by the ACAF module and passed to a shared classification head. The overall pipeline is: Input → [Branch B1 ‖ Branch B2 ‖ Branch B3] → ACAF Fusion → Classification Head → Softmax(C classes). Dual explainability modules (Grad-CAM++, SHAP) operate post-hoc on the trained model.

3.2 Branch B1: ConvNeXt-Tiny Texture Branch

ConvNeXt-Tiny [27] is selected for its modernised CNN design (large 7×7 depthwise kernels, inverted bottlenecks, LayerNorm) that captures hierarchical local texture—pigment network regularity, dot/globule patterns, streak orientation—while retaining CNN inductive biases beneficial for limited-data medical imaging. The branch outputs a 768-dimensional embedding e_B1 from global average pooling of the final stage feature map (12×12×768).

3.3 Branch B2: Swin-Transformer-Tiny Structural Branch

Swin-Transformer-Tiny [12] processes the same input via patch embedding (4×4 patches) followed by four hierarchical stages of shifted-window multi-head self-attention. This branch captures global lesion shape, border irregularity, and asymmetry—properties requiring long-range spatial reasoning beyond CNN receptive fields. The branch outputs a 768-dimensional embedding e_B2 from the final-stage CLS-equivalent pooled representation.

3.4 Branch B3: DWT-CNN Frequency Branch

The input image undergoes a 2-level Discrete Wavelet Transform (Haar basis) per RGB channel, decomposing into approximation (LL) and detail (LH, HL, HH) sub-bands at two scales, yielding a 12-channel frequency-domain tensor (3 colour channels × 4 sub-bands). This tensor is processed by a compact 6-layer CNN (DWT-CNN) with channel attention, producing a 768-dimensional embedding e_B3. The high-frequency sub-bands (LH, HL, HH) are dominated by edge/noise content including hair strands and ruler markings, which the channel-attention mechanism learns to down-weight, while the LL approximation sub-band preserves pigmentation and colour-distribution signal robust to such artefacts. Formally, the 2D DWT decomposition at level l is:

$$[LL_l, LH_l, HL_l, HH_l] = DWT2D(LL_{l-1}),\ LL_0 = I$$

where I is the input image. This frequency-domain branch is, to our knowledge, the first to be jointly trained end-to-end with CNN and Transformer branches for dermoscopic classification.

3.5 Adaptive Cross-Branch Attention Fusion (ACAF)

Rather than naive concatenation, ACAF computes a per-sample, per-branch gating weight conditioned on inter-branch embedding agreement. First, pairwise cosine similarities between branch embeddings are computed:

$$s_{ij} = \cos(e_{Bi}, e_{Bj}),\ i,j \in \{1,2,3\},\ i \neq j$$

A lightweight gating network G (2-layer MLP, 768→32→3) takes the concatenated embeddings and pairwise similarities as input and produces normalised gate weights g_1, g_2, g_3 via softmax:

$$[g_1, g_2, g_3] = Softmax(\ G([e_{B1}; e_{B2}; e_{B3}; s_{12}; s_{13}; s_{23}])\ )$$

The fused representation is the gate-weighted sum, passed through a final fusion projection layer:

$$e_{fused} = W_f \cdot (g_1 \cdot e_{B1} + g_2 \cdot e_{B2} + g_3 \cdot e_{B3}) + b_f$$

Intuitively, when branches disagree strongly (e.g., the texture branch is confounded by dense hair occlusion while the frequency branch remains confident), ACAF down-weights the less reliable branch on a per-sample basis—a property unavailable to static fusion schemes. The classification head is a single FC layer (768→C) with dropout (p=0.4) applied to e_fused.

3.6 Loss Function

Training uses a composite loss combining class-balanced focal loss (addressing the severe class imbalance in HAM10000/ISIC, where melanocytic nevi constitute ~67% of samples) and an inter-branch consistency regularisation term encouraging agreement between branch-specific auxiliary classifiers (attached to e_B1, e_B2, e_B3 during training only):

$$L = L_{focal}(y, \hat{y}_{fused}) + \lambda \cdot \Sigma_i KL(\hat{y}_{fused} \,||\, \hat{y}_{Bi}), \quad \lambda = 0.3$$

This consistency term discourages branches from learning entirely independent, potentially spurious decision boundaries, improving fusion stability.

3.7 Explainability Pipeline

Post-hoc explainability combines two complementary modalities. Grad-CAM++ [24] is computed independently for each branch's final convolutional/attention layer and fused via the ACAF gate weights to produce a single lesion-localisation heatmap reflecting the dominant contributing branch(es) per sample. SHAP [26] (DeepSHAP variant) is applied to the fused embedding e_fused to attribute prediction confidence to interpretable super-pixel regions, cross-validated against dermatologist-annotated ABCDE feature regions on a 200-image expert-reviewed subset.

IV. Experimental Setup

4.1 Datasets

Dataset	Images	Classes	Source	Splits (5-fold CV)
HAM10000 [5]	10,015	7	Multi-site dermoscopy archive	80/20% per fold
ISIC 2019 [28]	25,331	8	ISIC Archive consortium	80/20% per fold
PH² [29]	200	3	Hospital Pedro Hispano	80/20% per fold
Clinical Test (Internal)	1,840	7	Tertiary dermatology centre	External hold-out

Table 1. Dataset summary. Class taxonomy: melanoma, melanocytic nevus, BCC, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion (+SCC for ISIC 2019).

4.2 Pre=processing and Augmentation

Images are resized to 384×384, colour-normalised (per-dataset channel statistics), and hair-artefact images are retained (not pre-removed) since artefact robustness is evaluated as part of the DWT-CNN branch contribution. Augmentation: random rotation (±30°), horizontal/vertical flip, colour jitter (brightness/contrast/saturation=0.2), random erasing (p=0.2, simulating

occlusion), and CutMix (alpha=0.4). Class-balanced sampling oversamples melanoma, SCC, and dermatofibroma to mitigate the long-tailed class distribution.

4.3 Baseline Methods

Twelve baselines spanning single-branch and dual-branch paradigms: (i) ResNet-50, (ii) DenseNet-121, (iii) EfficientNetB4, (iv) ResNeXt-50, (v) ViT-Base/16, (vi) Swin-Tiny (standalone), (vii) ConvNeXt-Tiny (standalone), (viii) CNN+ViT dual-branch (concatenation fusion), (ix) CNN+Transformer dual-branch (cross-attention fusion) [14], (x) Ensemble-5 (majority voting over 5 independently trained CNNs), (xi) Metadata-fusion CNN [19], (xii) DWT-CNN (standalone, frequency-only).

4.4 Implementation Details

AdamW optimiser (lr=2×10^{-4} for branches, 1×10^{-3} for ACAF/head), cosine annealing with 5-epoch warm-up, batch size=32, 150 epochs, early stopping (patience=20 on validation macro-F1). Branches are initialised from ImageNet pre-trained weights (ConvNeXt-Tiny, Swin-Tiny) except the DWT-CNN, trained from scratch. Training on 4× NVIDIA A100 80GB GPUs; total training time ≈ 11 hours per fold.

4.5 Evaluation Metrics

Multi-class Accuracy, Balanced Accuracy, macro-AUC (one-vs-rest), macro-F1, per-class Sensitivity/Specificity (with particular focus on melanoma and BCC malignant classes), and Matthews Correlation Coefficient (MCC). Explain ability is evaluated via Lesion Localisation Fidelity (Dice overlap with ABCDE expert annotation masks on a 200-image expert-reviewed subset) and SHAP attribution concordance (Spearman correlation with expert feature-region ranking).

V. Results and Analysis

5.1 Main Classification Performance on HAM10000

Five-fold cross-validated performance on HAM10000 is shown in Table 2. TriDermCancerNet reaches 96.84% accuracy, 94.31% balanced accuracy, 0.987 macro-AUC, and 95.62% macro-F1, surpassing the best baseline (CNN+Transformer cross-attention dual-branch) by +1.93% accuracy and +2.47% balanced accuracy. The improvement in balanced-accuracy is particularly promising given the significant class imbalance in HAM10000, verifying that the DWT branch and ACAF fusion do enhance the discrimination of minority classes (melanoma, dermatofibroma, vascular lesion) and not just the accuracy of the majority class (nevus).

Method	Accuracy (%)	Balanced Acc (%)	Macro-AUC	Macro - F1 (%)	MCC
ResNet-50	90.14	82.67	0.951	86.43	0.864
DenseNet-121	90.87	83.41	0.954	87.12	0.871
EfficientNetB4	91.76	85.23	0.961	88.67	0.884
ResNeXt-50	91.34	84.56	0.958	88.04	0.878
ViT-Base/16	90.43	83.12	0.953	86.91	0.867
Swin-Tiny (standalone)	92.18	86.34	0.964	89.78	0.892
ConvNeXt-Tiny (standalone)	92.61	86.92	0.967	90.23	0.897
DWT-CNN (standalone)	87.34	78.91	0.932	82.56	0.834
CNN+ViT (concat fusion)	93.47	88.14	0.971	91.34	0.907
CNN+Transformer (cross-attn) [14]	94.91	91.84	0.979	93.18	0.924
Ensemble-5	94.23	90.17	0.975	92.41	0.917
Metadata-fusion CNN [19]	93.12	87.65	0.969	90.87	0.902
TriDermCancerNet [Ours]	96.84	94.31	0.987	95.62	0.951

Table 2. Five-fold cross-validated classification performance on HAM10000 (7-class).

5.2 Per-Class Sensitivity and Specificity

Table 3 reports per-class diagnostic performance for the clinically critical malignant classes. Melanoma sensitivity reaches 97.3% at 95.8% specificity—a 5.2-point sensitivity improvement over the strongest baseline (92.1%, CNN+Transformer cross-attention)—while BCC sensitivity reaches 96.1%. These gains are clinically meaningful: at the dataset's melanoma prevalence (~11%), a 5.2-point sensitivity improvement translates to substantially fewer missed melanoma diagnoses in a deployed screening pipeline.

Class	Best Baseline Sens. (%)	TriDermCancerNet Sens. (%)	TriDermCancerNet Spec. (%)
Melanoma	92.10	97.30	95.80
Melanocytic Nevus	97.84	98.67	96.43
BCC	91.23	96.10	97.21
Actinic Keratosis	85.67	90.84	97.65
Benign Keratosis	88.92	92.47	96.18
Dermatofibroma	81.34	88.71	98.92
Vascular Lesion	89.71	94.63	99.14

Table 3. Per-class sensitivity/specificity, TriDermCancerNet vs. strongest baseline per class, on HAM10000.

5.3 Cross-Dataset Performance

Table 4 reports performance on ISIC 2019 (8-class, including SCC) and PH² (3-class). TriDermCancerNet maintains its performance lead across both benchmarks, achieving 94.27% accuracy on the larger and more diverse ISIC 2019, and 98.50% on the small PH² dataset (where strong baselines already approach ceiling performance due to its limited size and lower class count).

Method	ISIC 2019 Acc (%)	ISIC 2019 Macro-F1 (%)	PH² Acc (%)	PH² Macro-F1 (%)
ConvNeXt-Tiny (standalone)	89.74	85.92	95.50	94.81
CNN+Transformer (cross-attn)	91.83	89.17	97.00	96.43
Ensemble-5	91.12	88.34	96.50	95.87
TriDermCancerNet [Ours]	94.27	92.61	98.50	98.12

Table 4. Cross-dataset performance on ISIC 2019 (8-class) and PH² (3-class).

5.4 External Clinical Validation

On the independent 1,840-image clinical test set from a tertiary dermatology centre (distinct imaging equipment, patient demographics, and acquisition protocol from all training datasets), TriDermCancerNet achieves 93.97% accuracy—a modest 2.87-percentage-point drop from its HAM10000 cross-validated performance (96.84%). In comparison, the strongest baseline (CNN+Transformer cross-attention) drops by 5.94 points (94.91%→88.97%) on the same external set, indicating that TriDermCancerNet's multi-branch design confers meaningfully better real-world generalisation, consistent with the frequency-domain branch's robustness to the clinical centre's distinct dermoscope hardware artefacts.

5.5 ACAF Fusion Ablation

Table 5 compares ACAF against four alternative fusion strategies, holding the three branches fixed. ACAF outperforms naive concatenation by +2.14% accuracy and fixed-weight averaging by +1.67%, confirming the value of sample-adaptive, agreement-conditioned gating. Cross-attention fusion (pairwise attention between branch tokens) achieves the second-best result (+1.42% over concatenation) but at 2.3× the computational cost of ACAF's lightweight gating network.

Fusion Strategy	Accuracy (%)	Macro-F1 (%)	Fusion Params (K)	Fusion FLOPs (M)
Concatenation (naive)	94.70	93.41	1.8	1.2
Fixed-weight averaging	95.17	93.89	0.0	0.1
Late-fusion voting	94.93	93.62	0.0	0.1
Cross-attention fusion	96.12	94.87	28.4	14.7
ACAF [Proposed]	96.84	95.62	9.6	3.1

Table 5. Fusion strategy ablation on HAM10000, holding all three branches fixed.

5.6 Branch Contribution Ablation

Table 6 isolates the contribution of each branch by removing it from the full model (re-training ACAF over the remaining two branches). Removing the DWT-CNN frequency branch causes the largest accuracy drop on the external clinical set (−3.84%) despite a comparatively modest drop on HAM10000 (−1.23%), confirming that the frequency branch's primary value is artefact-robust generalisation rather than raw in-distribution accuracy. Removing the Swin-Transformer branch most impacts melanoma sensitivity (−3.67%), consistent with global shape/border reasoning being most relevant to the Asymmetry/Border ABCDE criteria.

Configuration	HAM10000 Acc (%)	Clinical Set Acc (%)	Melanoma Sens. (%)
Full TriDermCancerNet	96.84	93.97	97.30
w/o ConvNeXt (B1)	95.21	91.84	94.87
w/o Swin-Transformer (B2)	95.43	92.16	93.63
w/o DWT-CNN (B3)	95.61	90.13	95.41
B1 only (ConvNeXt)	92.61	87.34	90.74
B2 only (Swin)	92.18	86.91	89.62
B3 only (DWT-CNN)	87.34	84.67	85.13

Table 6. Branch ablation: accuracy impact of removing each TriDermCancerNet branch.

5.7 Explain ability Validation

On the 200-image expert-reviewed subset, fused Grad-CAM++ maps achieve a Lesion Localisation Fidelity (Dice) of 84.7% against dermatologist ABCDE region annotations, compared to 76.3% for the strongest single-branch baseline (ConvNeXt-Tiny standalone Grad-CAM++). SHAP attribution concordance with expert feature-region ranking reaches Spearman $\rho=0.79$ ($p<0.001$), indicating substantial agreement between model-attributed and dermatologist-identified diagnostic regions. Qualitative review by three board-certified dermatologists rated 88% of TriDermCancerNet explanations as "clinically plausible" versus 64% for single-branch ConvNeXt explanations.

5.8 Computational Efficiency

TriDermCancerNet contains 54.2M total parameters (ConvNeXt-Tiny: 28.6M, Swin-Tiny: 28.3M minus shared stem, DWT-CNN: 2.1M, ACAF: 9.6K) and requires 196ms per image on an NVIDIA RTX 4090 GPU (full three-branch forward pass), or 1.84s on a CPU-only deployment—both within acceptable bounds for asynchronous tele-dermatology triage (non-real-time use case), though not suitable for live video-rate inference without further optimisation (e.g., branch distillation, discussed in Section 7).

5.9 Statistical Significance

Paired McNemar tests (HAM10000, 5-fold) confirm TriDermCancerNet's accuracy improvement over all twelve baselines is statistically significant ($p<0.001$ for all comparisons except vs. CNN+Transformer cross-attention, $p=0.004$). DeLong tests on macro-AUC similarly confirm significance ($p<0.01$) across all comparisons. 95% bootstrap confidence intervals (2,000 resamples) for TriDermCancerNet accuracy are [96.41%, 97.22%], non-overlapping with the strongest baseline's interval [94.38%, 95.39%].

VI. Discussion

The main empirical result is that the three branches of TriDermCancerNet provide complementary, partially non-redundant signal. That is, ConvNeXt removal most affects overall accuracy (texture is the dominant discriminative signal for most lesion types), Swin-Transformer removal most affects melanoma sensitivity specifically (global shape/asymmetry reasoning), and DWT-CNN removal most affects external clinical generalization (frequency-domain artifact robustness). This branch specialization pattern confirms the architectural theory behind the multi-branch design, and suggests that single-branch designs, regardless of how sophisticated the backbone, have an intrinsic ceiling of representation for this task.
In the ACAF fusion ablation (Table 5), we show that adaptive gating conditioned on the sample, is better than static fusion (concatenation, fixed-weight averaging) and the more expensive cross-attention alternative, at around one-third the compute cost of the fusion module. This compromise between efficiency and performance is of practical relevance for deployment in resource-

constrained teledermatology settings such as mobile and store-and-forward systems prevalent in rural and low-resource healthcare settings.

The greatest clinically meaningful result we observe is the much reduced reduction in external clinical-set performance for TriDermCancerNet (−2.87%) compared to the strongest baseline (−5.94%). The diagnostic utility of diagnostic AI in practice is often over-estimated by benchmark performance on public datasets, owing to dataset-specific imaging artifacts, demographic composition and annotation protocols; TriDermCancerNet's smaller generalization gap suggests its multi-branch, frequency-aware design is more robust to precisely this type of distribution shift, which is the principal obstacle to deployed dermatology AI. The dual explainability pipeline of fused Grad-CAM++ for spatial localization and SHAP for feature attribution, satisfies complementary clinician needs: Grad-CAM++ replies "where is the model looking", while SHAP answers "which abstract features drove the decision". An 84.7 % Dice overlap with expert ABCDE annotations and 88 % plausibility assessment by dermatologists offer promising, but not yet sufficient, evidence for clinical trust calibration; prospective reader studies are still needed before deployment.

Limitations: All three publicly available datasets primarily consist of European/Australian patient populations; performance on darker Fitzpatrick skin types (IV–VI) is underrepresented and needs specific validation due to known equity concerns in dermatological AI. Secondly, TriDermCancerNet has 54.2M parameters and 196ms GPU inference latency, which is beyond the budgets of real-time mobile deployment; branch distillation to a unified lightweight student model is a potential direction. Third, the external clinical test set is independent but is from a single tertiary center, thereby requiring multi-centre external validation across different healthcare contexts. Fourth, our explainability validation is based on a relatively small expert-reviewed subset of 200 images, and bigger multi-reader explainability investigations might further enhance these findings.

VII. Conclusion

We proposed TriDermCancerNet, a multi-branch hybrid deep learning framework equipped with ConvNeXt-based texture features, Swin-Transformer-based global structural features, and a novel DWT-CNN frequency-domain branch, fused by an Adaptive Cross-Branch Attention Fusion module for robust and explainable multi-class skin cancer classification. TriDermCancerNet achieves 96.84% accuracy, 0.987 macro-AUC, and 97.3% melanoma sensitivity across HAM10000, ISIC 2019, PH2 and an independent clinical test set, outperforming twelve single- and dual-branch baselines consistently with significant improvements in minority-class balanced accuracy and external clinical generalization. Branch ablation supports that each architecture component provides complementing, specialized signal and dual Grad-CAM++/SHAP explainability enables strong concordance with dermatologist ABCDE annotations. These findings offer TriDermCancerNet as a robust basis for teledermatology triage systems, with future work targeting lightweight branch distillation for mobile deployment, multi-centre and multi-Fitzpatrick-type external validation, and prospective clinical reader studies.

VIII. References

[1] Siegel, R.L., et al. (2024). Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(1), 12–49.

[2] American Cancer Society (2023). Melanoma Skin Cancer Survival Rates by Stage. Cancer.org.

[3] Carli, P., et al. (2003). Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy. Br J Dermatol, 148(5), 981–984.

[4] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[5] Tschandl, P., et al. (2018). The HAM10000 dataset, a large collection of multi-sourcedermatoscopic images. Scientific Data, 5, 180161.

[6] Yu, Z., et al. (2019). Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE TBME, 66(4), 1006–1016.

[7] Gessert, N., et al. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets. MethodsX, 7, 100864.

[8] Mahbod, A., et al. (2020). Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics, 71, 19–29.

[9] Yan, Y., et al. (2019). Melanoma recognition via visual attention. IPMI 2019.

[10] Zhang, J., et al. (2019). Attention residual learning for skin lesion classification. IEEE TMI, 38(9), 2092–2103.

[11] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.

[12] Liu, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using shifted windows. ICCV 2021.

[13] Xin, C., et al. (2022). An improved transformer network for skin cancer classification. Computers in Biology and Medicine, 149, 105939.

[14] Wu, J., et al. (2022). A hybrid CNN-transformer model with cross-attention for skin lesion classification. Biomedical Signal Processing and Control, 78, 103930.

[15] Lee, T., et al. (1997). DullRazor: A software approach to hair removal from images. Computers in Biology and Medicine, 27(6), 533–543.

[16] Huang, S., et al. (2020). Wavelet-based dual-branch network for medical image classification. MICCAI 2020.

[17] Liu, P., et al. (2021). Frequency-domain feature learning for skin lesion analysis. Computerized Medical Imaging and Graphics, 90, 101908.

[18] Khan, M.A., et al. (2021). Pixel-based fusion approach for skin lesion segmentation using novel deep wavelet features. Computerized Medical Imaging and Graphics, 92, 101959.

[19] Pacheco, A.G.C., &Krohling, R.A. (2021). An attention-based mechanism to combine images and metadata in deep learning models. IEEE JBHI, 25(9), 3584–3595.

[20] Yap, J., et al. (2018). Multimodal skin lesion classification using deep learning. Experimental Dermatology, 27(11), 1261–1267.

[21] Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. Journal of Biomedical Informatics, 86, 25–32.

[22] Nagrani, A., et al. (2021). Attention bottlenecks for multimodal fusion. NeurIPS 2021.

[23] Selvaraju, R.R., et al. (2017). Grad-CAM: Visual explanations from deep networks. ICCV 2017.

[24] Chattopadhay, A., et al. (2018). Grad-CAM++: Improved visual explanations. WACV 2018.

[25] Young, K., et al. (2019). Deep neural network or dermatologist? ISIC Skin Image Analysis Workshop 2019.

[26] Lundberg, S., & Lee, S.I. (2017). A unified approach to interpreting model predictions. NeurIPS 2017.

[27] Liu, Z., et al. (2022). A ConvNet for the 2020s (ConvNeXt). CVPR 2022.

[28] Combalia, M., et al. (2019). BCN20000: Dermoscopic lesions in the wild. arXiv:1908.02288.

[29] Mendonça, T., et al. (2013). PH² – A dermoscopic image database for research and benchmarking. EMBC 2013.