International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

Classification of Visually Similar Scalp Diseases using Deep Learning: A Hybrid CNN-VIT Approach with Cross-Attention Fusion

Research Scholar Ayushi Dixit, Dr. Brij Mohan Singh
Department of Computer Science, Quantum University, Roorkee, UK

Abstract- — Accurate automated diagnosis of visually similar scalp diseases represents one of the most challenging problems in clinical dermatology. Conditions such as Psoriasis, Seborrheic Dermatitis, Tinea Capitis, Alopecia Areata, Folliculitis, and Eczema share overlapping visual characteristics: including redness, scaling, and patchy hair loss, making misclassification clinically dangerous and common even among trained dermatologists. The global shortage of specialist dermatologists, particularly in rural and resource-limited settings in India, further amplifies the need for reliable automated diagnostic tools. This comprehensive research proposes ScalpViT, a novel hybrid deep learning architecture that combines a 16×16 Patch Vision Transformer (ViT) with a Convolutional Neural Network (CNN) backbone connected via a bidirectional cross-attention fusion module. The ViT branch processes the scalp image by dividing it into 256 non-overlapping 16×16-pixel patches, embedding each as a 768-dimensional token, and applying multi-head self-attention across the full token sequence to capture global spatial distribution and morphological patterns. Concurrently, the CNN branch extracts local texture details. The bidirectional cross-attention enables texture features to query spatial features and vice-versa, avoiding the pitfalls of simple feature concatenation. Trained on a meticulously curated multi-source dataset of approximately 7,000 dermoscopic and clinical scalp images drawn from DermNet NZ, ISIC 2018, HAM10000, and SD-198, ScalpViT achieves 94.3% accuracy, a macro F1-score of 0.93, and an AUC of 0.97. It significantly outperforms conventional baselines like ResNet-50 (83.1%), EfficientNet-B3 (87.4%), standard ViT-B/16 (90.8%), Swin-Tiny (91.2%), and DINOv2-B (93.5%). Furthermore, to bridge the interpretability gap for clinical deployment, ScalpViT utilizes GradCAM for CNN texture heatmapping and Attention Rollout for ViT patch mapping, delivering dual visual explainability to clinicians. The paper extensively details the methodology, dataset construction, architectural innovations, and clinical relevance for point-of-care mobile deployments.

Keywords— Vision Transformer (ViT), 16×16 Patch Tokenization, CNN-ViT Hybrid, Cross-Attention Fusion, Scalp Disease Classification, Psoriasis, Seborrheic Dermatitis, Medical Image Analysis, GradCAM, Explainable AI, EfficientNet-B3, DINOv2.

I. INTRODUCTION

Background and Motivation

Skin and scalp conditions are among the most widespread health concerns affecting people across the world. Studies estimate that dermatological problems affect more than one third of the global population at any given point in time, and scalp disorders in particular contribute significantly to this burden. Despite how common these conditions are, their diagnosis remains one of the most subjective and difficult tasks in clinical practice. Unlike many other medical conditions where a blood test or scan can provide a definitive answer, scalp diseases are diagnosed primarily through visual inspection, which introduces a heavy dependence on the experience and judgment of the examining clinician.

A patient who walks into a clinic with white flaky patches on the scalp could be suffering from Psoriasis, Seborrheic Dermatitis, Tinea Capitis, or something as benign as common dandruff. Each of these conditions requires a completely different line of treatment. Prescribing a steroid cream to someone who actually has a fungal infection will not help them and may make things worse. Treating someone for dandruff when they actually have scalp Psoriasis will leave the underlying autoimmune condition unaddressed. The visual overlap between these conditions is so significant that even experienced dermatologists sometimes disagree on the correct diagnosis without additional tests.

In India, there are roughly 10,000 certified dermatologists serving a population of approximately 1.4 billion people. This works out to fewer than one specialist for every 100,000 citizens, which is far below what the World Health Organization considers adequate. The majority of these specialists are concentrated in large cities and private hospitals, while rural communities, which account for nearly 65 percent

of the Indian population, have very limited access to specialist care. Patients living in remote areas often have to travel for hours and spend money they cannot easily afford just to receive a diagnosis. Many of them end up relying on a general practitioner or a local pharmacist who, without specialised training in dermatology, may not be equipped to distinguish between visually similar scalp conditions. This leads to treatment delays, repeated consultations, unnecessary medication, and in some cases, progression of a condition that could have been caught early.

An intelligent computer-based system that can look at an image of a patient's scalp and reliably identify what condition they might be suffering from would be enormously useful in this context. Such a tool would not replace the dermatologist but could serve as a first-line screening aid, helping general practitioners and community health workers make more informed decisions about when and where to refer patients.

II. PROBLEM STATEMENT AND CHALLENGES

Given a scalp image I belonging to the space R^(H×W×3), which may be a dermoscopic image or a standard clinical photograph, the goal is to design and train a deep learning model f such that f(I) = y, where y is one of six disease classes: Psoriasis, Seborrheic Dermatitis, Tinea Capitis, Alopecia Areata, Folliculitis, or Eczema. This poses severe challenges:

1) Visual Ambiguity: The most immediate difficulty is that the six target conditions look very similar to each other, especially in their early or mild stages. Psoriasis and Seborrheic Dermatitis both produce scaling and redness on the scalp. Tinea Capitis and Alopecia Areata both produce patches of hair loss. Folliculitis and Eczema can both present as red, irritated skin. A computer vision model must rely entirely on subtle differences in colour, texture, shape, and spatial distribution that may be extremely hard to separate even for an experienced eye.

2) Dataset Scarcity: Deep learning models in general require large numbers of labelled examples. For general skin lesion classification, several large public datasets exist (e.g., HAM10000, ISIC). But these datasets focus primarily on lesions of the body rather than the scalp. Building a dedicated scalp disease dataset from scratch is expensive and time-consuming. Any model developed must therefore learn from a relatively small dataset without overfitting.

3) Architectural Limitation: CNNs are good at picking up local texture features, such as the granularity of scaling, the appearance of individual pustules, or the colour pattern within a small region. But they struggle to capture the global spatial structure of the disease, such as whether the scaling is evenly spread across the scalp or concentrated in specific zones. Vision Transformers (ViTs) are naturally good at modelling these global relationships because self-attention operates across all patches simultaneously, but they tend to be less sensitive to fine-grained local texture. Designing an architecture that genuinely combines local and global feature extraction is a non-trivial engineering challenge.

4) Class Imbalance: Rarer conditions like Folliculitis and Tinea Capitis have far fewer images available compared to Psoriasis and Seborrheic Dermatitis. When a model is trained on an imbalanced dataset, it naturally learns to favour the majority classes. Addressing this requires deliberate choices in the loss function, sampling strategy, and data augmentation approach.

5) Interpretability Gap: Finally, even a model that achieves high classification accuracy cannot automatically be trusted in a clinical setting. A clinician receiving an AI diagnosis needs to understand the basis for that diagnosis. Providing visual explanations (heatmaps/attention maps) is not an optional extra; it is a requirement for responsible clinical deployment.

Research Objectives

- Design and implement a 16×16 patch ViT encoder for scalp image tokenization and establish a baseline classification performance.
- Develop a CNN and ViT hybrid architecture (ScalpViT) with bidirectional cross-attention fusion of local and global image features.
- Construct and preprocess a multi-source scalp disease dataset with class-balanced augmentation strategies.
- Train and evaluate ScalpViT against CNN and ViT baselines using standard metrics including Accuracy, macro F1, AUC, and Cohen's Kappa.
- Apply GradCAM and Attention Rollout to produce dual visual explanations for clinical interpretability.
- Conduct an ablation study to quantify the individual contribution of each architectural component.

II. LITERATURE REVIEW

1. Traditional Machine Learning Approaches

Before deep learning became the dominant approach, automated dermatology systems relied on manually extracting handcrafted image features. Color histograms captured the

distribution of red, green, and blue values. Local Binary Patterns (LBP) captured local texture patterns. Gabor wavelets responded to edges and textures at particular orientations and scales. SIFT detected distinctive local features like corners and blobs.

Once extracted, these features were fed into classical statistical classifiers like Support Vector Machines (SVMs), Random Forests, or k-Nearest Neighbours. While these systems achieved reasonable results on simple binary tasks (e.g., distinguishing melanoma from benign skin lesions), they lacked adaptability. The handcrafted features depended entirely on the designer's intuitions and struggled to capture the full visual complexity and high variability of scalp diseases.

2. CNN-Based Deep Learning for Dermatology

The arrival of Convolutional Neural Networks (CNNs) revolutionized medical image analysis. Instead of relying on manual feature design, CNNs learn relevant features directly from data. Architectures like AlexNet, VGGNet, ResNet, InceptionNet, and EfficientNet progressively raised the performance ceiling.

Transfer learning became critical: a CNN pretrained on a massive dataset of natural images (e.g., ImageNet) could be fine-tuned on a smaller medical dataset, achieving remarkable results. In 2017, a landmark paper by Esteva et al. demonstrated that a CNN could classify skin cancer from clinical images as accurately as board-certified dermatologists. Following this, Zhang et al. (2022) applied CNNs specifically to scalp diseases, distinguishing Scalp Psoriasis and Seborrheic Dermatitis from dermoscopic images with 96.1% sensitivity and an AUC of 0.922.

Despite this progress, CNNs possess an inherent limitation: local receptive fields. A CNN processes an image by looking at small local regions through its convolutional filters. To influence a pixel on the opposite side of the image, the signal must travel through many layers of convolutions. For scalp conditions where the spatial distribution of the disease matters as much as the local texture, this is a significant architectural weakness.

3. Vision Transformer (ViT) Architectures

Originating in natural language processing, the Transformer architecture utilizes self-attention to process entire sequences simultaneously. In 2020, Dosovitskiy et al. introduced the Vision Transformer (ViT), dividing an image into a grid of non-overlapping square patches (e.g., 16×16 pixels), flattening each patch, and applying a standard Transformer to this sequence of patch tokens.

In a ViT, every patch can directly attend to every other patch in the very first layer. The global context is available immediately. The Swin Transformer (Liu et al., 2021) refined this by applying self-attention within shifting local windows, restoring some hierarchical feature extraction. More recently, DINOv2 (Oquab et al., 2023) demonstrated that self-supervised pretraining on massive unlabelled datasets produces remarkably robust visual features, achieving state-of-the-art results on 31-class skin disease classification (Mohan et al., 2024). However, pure ViT models often demand large training datasets and lack the CNN's inductive bias for translation invariance, frequently underperforming in fine-grained texture discrimination on small datasets.

4. CNN-Transformer Hybrid Models

Recognizing the complementary strengths of CNNs (local texture) and ViTs (global structure), researchers have explored hybrid models. YoTransViT (Saha et al., 2024) merged CNN and Transformer features, utilizing image segmentation preprocessing. Empirical feature selection studies have verified that CNN-based backbones consistently learn local texture patterns, while ViT-based backbones focus on global spatial structure.

A major limitation of existing hybrid models is their reliance on simple feature concatenation for fusion. Concatenation treats the two sets of features as additive information with no explicit interaction, forcing the classifier to infer relationships from scratch. ScalpViT addresses this gap by utilizing bidirectional cross-attention, allowing CNN features to explicitly query relevant ViT features, and vice versa.

5. Multimodal AI and Explainability

The future of clinical dermatology lies in multimodal AI. Systems like SkinGPT-4 (Zhou et al., 2024) align a pretrained ViT with a large language model (LLaMA) to generate diagnostic reports from images and clinical notes. PanDerm (Yan et al., 2025) pushed this further, training a vision foundation model on 2 million diverse dermatology images across four imaging modalities.

Parallel to architectural advancements is the growing requirement for explainable AI (XAI). Classifiers must provide interpretable reasoning. While some CNN studies employ GradCAM, few systems provide comprehensive explainability for hybrid architectures. ScalpViT fills this gap by implementing dual explainability: GradCAM for the CNN branch and Attention Rollout for the ViT branch.

III. BACKGROUND THEORY

1. Self-Attention Mechanism

The self-attention mechanism computes pairwise relationships between all elements of an input sequence. Given an input sequence X ∈ R^(N×D), three linear projections produce queries Q, keys K, and values V:

$$Q = XW\_Q, K = XW\_K, V = XW\_V$$

The attention output is computed as the weighted sum of values, where weights are derived from the scaled dot-product of queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d\_k}) \cdot V$$

For the 16×16 patch configuration with a 256×256 input, the sequence length is N = 256 patch tokens + 1 [CLS] token = 257, which is computationally tractable on modern hardware with complexity $O(N^2D)$.

2. 16×16 Patch Vision Transformer

The input image I ∈ R^(H×W×C) is divided into $N = HW/P^2$ non-overlapping patches of size P×P. For our configuration: H = W = 256, P = 16, giving N = 256 patches. Each patch is a 16×16×3 = 768-dimensional vector after flattening the RGB channels.

Each flattened patch x_i is projected to a D-dimensional embedding via a learnable matrix E. A learnable [CLS] classification token is prepended, and learnable positional encodings E_pos are added. The sequence is processed by L = 12 stacked encoder blocks, each with Multi-Head Self-Attention (MSA) and a Feed-Forward Network (FFN), connected via residual connections and LayerNorm. The final [CLS] token is used as the image-level representation.

3. Cross-Attention Fusion

Cross-attention differs from self-attention in that queries come from one modality while keys and values come from another. In ScalpViT, projected CNN feature tokens F_cnn serve as queries while ViT patch tokens F_vit serve as keys and values: $\text{CrossAttn}(F\_cnn, F\_vit) = \text{softmax}(F\_cnn\, W\_Q\, (F\_vit\, W\_K)^T / \sqrt{d\_k}) \cdot F\_vit\, W\_V$

This allows CNN features to attend to the most relevant ViT tokens, enriching local texture representations with global spatial context. The reverse operation (ViT attending to CNN) runs in parallel, and the two outputs are concatenated and normalised to produce the fused representation.

4. Explainability Methods

1) GradCAM: Gradient-weighted Class Activation Mapping traces the gradient of the predicted class score backwards through the CNN to the convolutional feature maps. It computes how much a small change in a specific location of the feature map would change the final class prediction, producing a spatial heatmap of clinical importance.

2) Attention Rollout: For the ViT branch, Attention Rollout recursively multiplies attention weight matrices backwards through all 12 transformer encoder blocks. This traces which input patches the final [CLS] token attended to, producing a 16×16 attention grid reflecting the global spatial patterns the network prioritized.

IV. METHODOLOGY AND SYSTEM DESIGN

1. Dataset Description

A dedicated large-scale scalp disease dataset does not currently exist. We therefore aggregated scalp-relevant images from four public dermatology datasets, filtering and relabelling for the six target disease classes. DermNet NZ provided ~3,200 clinical photos. ISIC 2018 provided ~1,800 dermoscopic images. HAM10000 provided ~900 dermoscopic images. SD-198 provided ~1,100 clinical photos. After deduplication and relabelling, the aggregated dataset contained approximately 7,000 labelled scalp images. Class distribution is inherently imbalanced: Psoriasis and Seborrheic Dermatitis constitute approximately 55% of the dataset, while Folliculitis and Tinea Capitis each contribute under 10%.

2. Data Preprocessing and Augmentation

To address dataset scarcity and class imbalance, robust data augmentation techniques were deployed. Standard geometric transforms included Random Horizontal Flip (p=0.5) and Random Rotation (±20°). Colour Jitter (brightness, contrast, saturation) normalized lighting variations. Advanced augmentations included CutOut (randomly masking 32×32 pixel regions) to promote robust feature learning, MixUp (creating virtual training examples by interpolating pairs of images and labels with α=0.2) to smooth decision boundaries, and RandAugment for automated augmentation search. Additionally, minority classes were oversampled using synthetic images generated via Stable Diffusion XL.

3. ScalpViT Architecture Modules

1) Module 1 - Dual-Stream Feature Extractor: The CNN branch utilizes EfficientNet-B3. When an image (256×256) passes through the CNN, the final convolutional stage produces 384 feature maps of size 16×16. These are flattened into a sequence of 256 tokens, each with 384 dimensions. The ViT branch

utilizes ViT-B/16. The image is split into 256 patches of 16×16 pixels. Each patch is flattened, projected to 768 dimensions, combined with positional encodings and a [CLS] token, and passed through 12 Transformer encoder blocks. This results in 256 patch tokens of 768 dimensions capturing global spatial structure.

2) Module 2 - Cross-Attention Fusion: The 384-dimensional CNN feature vectors are projected via a learnable linear layer to 768 dimensions to match the ViT space. Bidirectional cross-attention is then computed. In the CNN-to-ViT attention, CNN tokens query ViT tokens, allowing local texture to integrate global spatial patterns. Simultaneously, the ViT-to-CNN attention allows global patterns to query local textures. The outputs ([B, 256, 768] each) are concatenated along the feature dimension, producing a [B, 256, 1536] tensor. This is mean-pooled across the 256 spatial positions to yield a single 1536-dimensional image-level representation, followed by batch normalization and dropout (0.1).

3) Module 3 - Classification Head: The classification head is an MLP implementing gradual dimensionality reduction: FC(1536 → 512), GELU activation, Dropout(0.3), FC(512 → 128), GELU activation, and FC(128 → 6). A final softmax function converts the logits into a probability distribution over the six disease classes.

4) Module 4 - Explainability Layer: Applied post-hoc, GradCAM computes gradients relative to the EfficientNet-B3 final convolutional layer, producing a continuous spatial heatmap. Attention Rollout propagates attention weights recursively across all 12 ViT layers to produce a 16×16 attention grid for the 256 input patches. Both maps are bilinearly upscaled to 256×256 and overlaid on the original image.

4. Training Strategy

Training follows a three-stage schedule to prevent catastrophic forgetting of pretrained weights. Stage 1 (Backbone freeze, 5 epochs): Only the fusion module and classification head are trained at lr = 1e-3. Stage 2 (Partial unfreeze, 15 epochs): The last 4 ViT blocks and last 2 CNN stages are unfrozen with layer-wise learning rate decay (factor 0.8) and base lr = 1e-4 with cosine annealing. Stage 3 (Full fine-tune, 10 epochs): All parameters are trained at lr = 1e-5. The composite loss function is $L_{total} = 0.5 \cdot L_{focal} + 0.3 \cdot L_{CE} + 0.2 \cdot L_{smooth}$, addressing class imbalance (Focal loss, γ=2.0), stabilizing gradients (Cross-Entropy), and preventing overconfidence (Label Smoothing).

V. EXPERIMENTS AND RESULTS

1. Experimental Setup

All experiments were conducted using PyTorch 2.1 on an NVIDIA A100 40GB GPU. The 7,000-image dataset was split into training (80%), validation (10%), and testing (10%) sets using stratified sampling. During development, 5-fold cross-validation was performed on the train/val split. The held-out test set (700 images) was evaluated exactly once to prevent data leakage.

2. Baseline Comparison

Six models were evaluated under identical conditions. The CNN baseline ResNet-50 achieved an Accuracy of 83.1% and Macro F1 of 0.810. EfficientNet-B3 (CNN) improved this to 87.4% and 0.860. The pure ViT-B/16 achieved 90.8% and 0.890. Swin-Tiny (Hierarchical ViT) reached 91.2% and 0.900. The self-supervised DINOv2-B achieved 93.5% and 0.920. The proposed ScalpViT achieved the highest performance across all metrics: 94.3% Accuracy, 0.930 Macro F1, 0.972 AUC, and 0.932 Cohen's κ. This demonstrates that combining texture and spatial features explicitly yields superior classification.

3. Ablation Study

An ablation study quantified each component's contribution. EfficientNet-B3 (CNN only) served as the 87.4% baseline. ViT-B/16 only added +3.4% (90.8%). A hybrid using simple concatenation fusion improved performance to 92.0% (+4.6% over baseline). Replacing concatenation with bidirectional cross-attention fusion yielded 93.6% (+6.2% over baseline, +1.6% over concatenation). Adding the composite loss function (focal + label smoothing) resulted in the final 94.3% accuracy (+6.9% total improvement). This empirically validates the superiority of learned cross-attention fusion over concatenation.

4. Confusion Matrix Analysis

The normalized confusion matrix revealed that the most frequent error (approx. 3% in both directions) was confusing Seborrheic Dermatitis with Psoriasis. This is clinically expected, as both present with scaling and diffuse redness; Psoriasis typically features thick, silvery scales, while Seborrheic Dermatitis presents with yellowish, greasy scales, but borderline cases challenge even expert dermatologists. The second most common error (2%) was between Tinea Capitis and Alopecia Areata, both presenting as hair loss, highlighting ambiguity in mild clinical presentations. ScalpViT excelled at isolating distinct classes like Folliculitis and Eczema.

5. Qualitative Visualisation

Applying the explainability layer to a Psoriasis test image revealed that the 16×16 Attention Rollout grid strongly prioritized patches containing thick silvery scaling. Similarly, the GradCAM overlay produced a high-intensity heat signature precisely over the scaly plaque, while ignoring background healthy hair and scalp tissue. This dual verification confirms that the model's reasoning relies on pathognomonic clinical signs rather than spurious correlations or background artifacts.

VI. DISCUSSION

1. Interpretation of Results and Clinical Relevance

ScalpViT's 94.3% accuracy represents a 6.9 percentage point improvement over a pure CNN (EfficientNet-B3). In a clinical setting processing 100 scalp patients daily, this equates to 7 additional correct diagnoses per day. Furthermore, distinguishing between six visually ambiguous conditions is a significantly harder information-theoretic problem than binary classification. ScalpViT's error patterns mirror clinical realities, struggling only where human clinicians experience genuine ambiguity.

The introduction of dual explainability is vital for clinical trust and regulatory compliance. Providing both an attention-based patch explanation and a gradient-based spatial heatmap allows a clinician to verify the diagnosis visually. If both metrics align over the affected lesion, trust in the AI recommendation increases dramatically, enabling the tool to function safely as a point-of-care screening aid for general practitioners in rural communities.

2. Comparison with Prior Work

While Zhang et al. (2022) achieved 96.1% sensitivity on a binary task (Psoriasis vs. Seborrheic Dermatitis), ScalpViT tackles a 6-class problem with severe inter-class similarity. Compared to Mohan et al. (2024), who achieved 96.48% on a 31-class skin disease task using DINOv2, ScalpViT operates in a much narrower visual domain where disease classes exhibit extreme overlap (e.g., exclusively scalp-based erythema/scaling). The ablation results confirm that ScalpViT's cross-attention fusion provides a direct and cost-free performance boost over the concatenation approaches utilized by earlier hybrid models.

VII. CONCLUSION AND FUTURE WORK

This research proposed ScalpViT, a novel hybrid deep learning architecture designed to classify six visually similar scalp diseases. By extracting local textures via an EfficientNet-B3 CNN and global spatial configurations via a 16×16 Patch ViT, and intelligently fusing these representations via bidirectional cross-attention, the model achieved an accuracy of 94.3% on a diverse dataset. The inclusion of GradCAM and Attention Rollout ensures transparent, interpretable decision-making, moving the system from a "black-box" classifier to a viable clinical aid.

Future research will focus on six pivotal areas. First, Multimodal LLM Integration (e.g., aligning ScalpViT with LLaMA-3) to generate comprehensive natural language diagnostic reports based on the image and textual clinical symptoms. Second, Federated Learning across hospital networks to expand the training dataset diversity without compromising patient privacy. Third, Mobile Deployment via model quantization (e.g., mapping to a quantized DINOv2-Small backbone) to allow offline inference on rural clinic smartphones. Fourth, Trichoscopy Specialisation, fine-tuning the model exclusively on high-magnification dermoscopic imaging. Fifth, Severity Estimation through regression heads to predict clinical indices like PASI or SALT. Finally, robust Fairness and Bias Evaluation across all Fitzpatrick skin types (I-VI) to guarantee equitable diagnostic accuracy for diverse patient populations, particularly those with darker skin tones where erythema presents atypically.

REFERENCES

1. M. Roy and A. T. Protity, "Hair and scalp disease detection using machine learning and deep learning techniques," 2023.
2. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
3. Y. Zhang, K. Shen, J. Han, G. Yu, and Y. Wang, "A deep learning-based approach toward differentiating scalp psoriasis and seborrheic dermatitis from dermoscopic images," Front. Med., vol. 9, p. 965423, 2022.
4. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019.
5. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
6. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. ICCV, 2021.
7. M. Oquab et al., "DINOv2: Learning robust visual features without supervision," TMLR, 2023.
8. J. Mohan et al., "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable AI," Comput. Biol. Med., vol. 190, 2024.

9. D. K. Saha et al., “YoTransViT: A transformer and CNN method for predicting and classifying skin diseases,” Inform. Med. Unlocked, vol. 47, 2024.
10. J. Zhou et al., “Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4,” Nat. Commun., vol. 15, 2024.
11. S. Yan et al., “A multimodal vision foundation model for clinical dermatology (PanDerm),” Nat. Med., 2025.
12. A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017.
13. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in ICCV, 2017.
14. S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," in ACL, 2020.