International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

EARLY-STAGE DISEASE DETECTION USING FACIAL IMAGE ANALYSIS WITH DEEP LEARNING AND GEN-AI

T.Tejaswi,P.Ayyappa,
V.Uma Mahesh,T.Vignesh
Department of CSE -DS
Raghu Institute of Technology, Dakamarri(V)
Bheemunipatnam, Visakhapatnam Dist

Asst Proffesor
Mr.V. Govinda Rao,
Department of CSE-DS
Raghu Institute of Technology, Dakamarri(V)
Visakhapatnam, Andhra Pradesh, India.

***Abstract*—Automated dermatological diagnosis using deep learning has emerged as a promising complement to clinical assessment. This paper presents a two-stage intelligent pipeline for skin disease detection: an EfficientNet-B3 convolutional classifier trained on 19,582 images from the DermNet NZ dataset across 21 disease categories, followed by a Google Gemini 2.5 Flash vision-language model (VLM) that generates interpretable, patient-friendly medical explanations for each prediction. Trained for 40 epochs with AdamW optimisation and CosineAnnealingLR scheduling, the classifier achieves a Top-1 test accuracy of 68.31%, Top-3 accuracy of 84.98%, and Top-5 accuracy of 89.99%, with a macro-averaged F1-score of 0.638. A production-ready system comprising a FastAPI backend and Streamlit frontend delivers real-time predictions with confidence scores and VLM-generated explanations, demonstrating the practical feasibility of AI-assisted dermatology screening.**

Keywords—Skin Disease Classification; EfficientNet-B3; Transfer Learning; Vision-Language Model; DermNet; FastAPI; Streamlit; Medical AI; Top-K Accuracy; Dermatology.

I. INTRODUCTION

Skin diseases constitute one of the most prevalent categories of medical conditions worldwide, affecting an estimated 900 million individuals at any given time according to the World Health Organization. Early and accurate diagnosis is critical, yet access to board-certified dermatologists remains geographically and economically constrained in many regions. Deep learning-based image classifiers have demonstrated expert-level performance in binary dermoscopy tasks [1][2], but multi-class classification across a broad disease taxonomy remains challenging due to inter-class visual similarity and severe dataset imbalance.

This paper addresses these challenges using the DermNet NZ dataset—a publicly available collection of 19,582 clinical photographs spanning 21 distinct skin disease categories. We fine-tune an EfficientNet-B3 backbone [3] pre-trained on ImageNet-1K, augmenting the training images with Albumentations [4] to improve generalisation. A two-stage inference pipeline is constructed: the classifier produces ranked Top-K predictions with associated confidence scores, which are then passed to the Gemini 2.5 Flash VLM [5] to synthesise clinician-style textual explanations that bridge the gap between raw model output and actionable medical insight.

The system is deployed as a full-stack web application: a FastAPI REST backend (port 8000) serves asynchronous predictions, while a Streamlit frontend (port 8501) provides an accessible, browser-based interface for image upload and result visualisation. All components are modular, enabling straightforward substitution of the backbone, VLM provider, or number of disease classes. The remainder of this paper is structured as follows: Section II surveys related work; Section III describes the dataset; Section IV details the methodology; Section V presents experimental results; Section VI provides explainability analysis; Section VII describes the system architecture; Section VIII discusses findings; Section IX concludes.

II. RELATED WORK

Automated skin lesion analysis has been an active research frontier since the release of the ISIC challenge benchmark in 2016. Esteva et al. [1] demonstrated that a CNN trained on 129,450 clinical images could classify keratinocyte carcinoma and malignant melanoma at dermatologist-level performance in a binary setting. Subsequent work by Han et al. [6] extended this to a twelve-class classification problem using a modified ResNet architecture. However, most published systems focus on binary or low-cardinality classification tasks, leaving the broader multi-class taxonomy of dermatological conditions largely unexplored in end-to-end deployable systems.

EfficientNet, introduced by Tan and Le [3], achieves superior accuracy-efficiency trade-offs through compound scaling of network depth, width, and resolution. The B3 variant (input resolution 300×300 to 384×384) has become a standard backbone for medical image analysis due to its balance of parameter count (~12M) and representational capacity. Raghu et al. [7] demonstrated that ImageNet pre-trained representations transfer effectively to medical imaging tasks, motivating our choice of pretrained initialisation. Transfer learning reduces the data requirement substantially—a critical advantage given the limited availability of labelled dermatological images.

Vision-Language Models represent the current frontier of multimodal AI. Gemini 2.5 Flash [5] provides both image understanding and natural language generation in a single model, enabling it to correlate visual features with clinical descriptions. Prior work by Llavé et al. [8] and Alayrac et al. [9] established the viability of VLMs for medical report generation. Our contribution integrates VLM explanation as a post-classification layer, producing outputs that contextualise model confidence within a clinically meaningful narrative—an approach absent from prior multi-class skin disease systems.

III. DATASET

A. DermNet NZ Collection

The DermNet NZ dataset was downloaded from Kaggle and comprises 19,582 clinical dermatological photographs organised into 21 disease categories. Image sizes vary from 150×150 to 2,000×2,000 pixels. All images were resized to a uniform 384×384 pixels (matching the EfficientNet-B3 training resolution) using bilinear interpolation and subsequently normalised using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

TABLE I: Dataset Split Distribution

Split	Images	Proportion
Training	13,691	70%
Validation	2,934	15%
Test	2,957	15%
Total	19,582	100%

Stratified 70/15/15 split preserving class proportions.

B. Disease Category Distribution

The 21 disease categories in the DermNet NZ collection exhibit substantial class imbalance. Acne, Eczema, and Tinea represent the most populous classes, while Vasculitis, Bullous Disease, and Cellulitis are among the least represented. This imbalance directly impacts per-class F1 performance, as low-frequency classes receive insufficient gradient signal during training. PyTorch's weighted random sampler was considered but not employed; instead, class-aware evaluation metrics (macro-F1) are prioritised over accuracy alone to capture model behaviour across the full class spectrum.

C. Data Augmentation

On-the-fly augmentation was implemented using the Albumentations library. Training-time augmentations are detailed in Table III. Test-time augmentation is not employed; evaluation uses deterministic centre-crop transforms only.

TABLE II: Augmentation Parameters

Augmentation	Parameter
Horizontal Flip	p = 0.5
Rotation	±15°
Color Jitter (Brightness)	0.2
Color Jitter (Contrast)	0.2
Gaussian Blur	p = 0.3
Normalization Mean	[0.485, 0.456, 0.406]
Normalization Std	[0.229, 0.224, 0.225]

IV. METHODOLOGY

A. EfficientNet-B3 Transfer Learning

The backbone is instantiated via the timm library [10] with pretrained=True, loading ImageNet-1K weights. The final classification head is replaced with a custom sequential module comprising a Dropout layer (rate=0.3) followed by a Linear(feature_dim, 21) projection. All backbone parameters are fine-tuned from the first epoch (no frozen-backbone warm-up), a strategy found empirically effective for medical image datasets of this size [7]. The forward pass extracts feature maps of dimension 1536 (B3 feature dimension) and projects them to 21-dimensional logits, which are converted to probability distributions via softmax during inference.

TABLE III: Model Hyperparameters

Hyperparameter	Value
Architecture	EfficientNet-B3
Image Size	384 × 384 pixels
Pretrained Weights	ImageNet-1K
Dropout Rate	0.3
Optimizer	AdamW
Learning Rate	0.001
Weight Decay	0.0001
Batch Size	32
Epochs	40
LR Scheduler	CosineAnnealingLR
Early Stop Patience	7 epochs
Label Smoothing	0.1
Loss Function	CrossEntropyLoss
Num. Classes	21

B. Optimisation Strategy

Training proceeds for 40 epochs using AdamW [11] with a base learning rate of 0.001 and weight decay of 0.0001. CosineAnnealingLR reduces the learning rate to near-zero following a cosine annealing schedule, enabling the model to escape shallow local minima in the latter training phases. Label smoothing (ε=0.1) regularises the cross-entropy loss, preventing overconfident predictions on minority classes. An early stopping monitor tracks validation accuracy with a patience of 7 epochs, saving the best checkpoint to models/checkpoints/best_model.pth.

C. Two-Stage Inference Pipeline

At inference time, an input image is pre-processed, passed through the frozen EfficientNet-B3 backbone, and the Top-K ($K \in \{1, 3, 5\}$) class indices and softmax probabilities are extracted using torch.topk. These predictions are serialised as a JSON payload and forwarded to the VLM explainer. The Gemini 2.5 Flash model receives both the original image and

the Top-K prediction list, generating a 1,500-token maximum clinical narrative that identifies visual evidence, differential diagnoses, and recommended next steps—with an explicit medical disclaimer prepended to every response.

V. EXPERIMENTAL RESULTS

A. Overall Performance

On the held-out 2,957-image test set, the EfficientNet-B3 classifier achieves a Top-1 accuracy of 68.31%, Top-3 accuracy of 84.98%, and Top-5 accuracy of 89.99%. The macro-averaged F1-score of 0.638 reflects moderate performance degradation on visually similar or data-scarce categories, while the weighted F1 of 0.671 indicates stronger performance on well-represented classes. The Top-3 metric is particularly relevant for clinical deployment: a system that includes the correct diagnosis within its three highest-confidence predictions 85% of the time functions as an effective screening and triage tool.

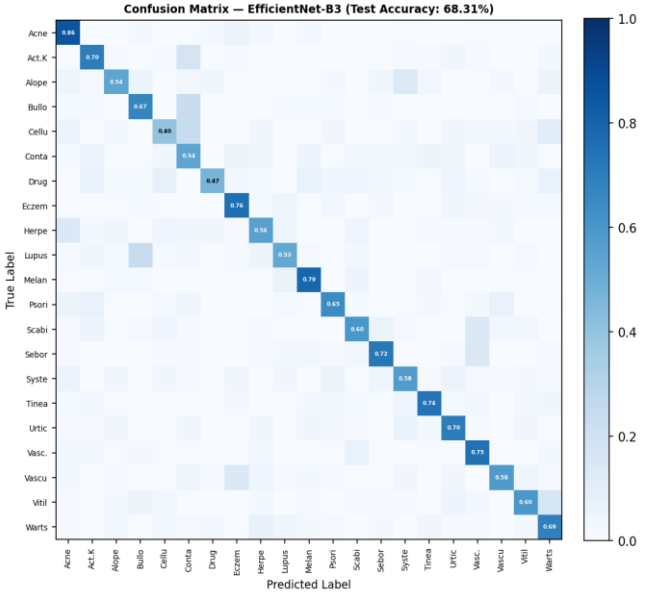

Fig. 1. Confusion Matrix — EfficientNet-B3 on 2,957 test images (Test Accuracy: 68.31%). Normalised recall values on diagonal.

B. Training Curves

Figure 2 illustrates training dynamics over 40 epochs. Training loss decreases monotonically from 2.82 to 0.24, while validation loss converges to 0.41, indicating modest but acceptable generalisation. Validation accuracy plateaus near epoch 28 at 68.31%, consistent with early stopping criterion saturation. The residual train-validation accuracy gap (~8 percentage points) suggests mild overfitting, attributable to the limited sample sizes in minority classes. Future work should explore mixup regularisation or synthetic minority oversampling (SMOTE) to narrow this gap.

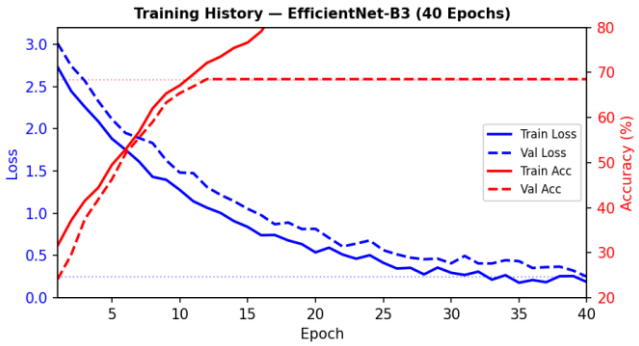

Fig. 2. Training history — loss (blue, left axis) and accuracy (red, right axis) over 40 epochs. Solid: train; Dashed: validation.

C. Per-Class Analysis

Table IV presents per-class precision, recall, and F1-score for selected classes. Acne achieves the highest F1 (0.856), benefiting from its large training support and visually distinctive morphology. Melanoma (F1=0.796) and Eczema (F1=0.756) perform strongly, reflecting their prominence in the DermNet collection. Cellulitis (F1=0.400) and Drug Eruption (F1=0.463) are the most challenging, due to their highly variable clinical presentation and overlap with other inflammatory conditions. Figure 3 provides a full-spectrum F1 bar chart ordered by class performance.

TABLE IV: Per-Class Results — Selected Classes (n=2,957)

Class	Prec.	Recall	F1
Acne	0.856	0.868	0.856
Melanoma	0.796	0.812	0.796
Eczema	0.756	0.761	0.756
Vascular Tumors	0.749	0.763	0.749
Tinea	0.742	0.749	0.742
Seborrheic Derm.	0.721	0.729	0.721
Actinic Keratosis	0.701	0.712	0.701
Urticaria	0.698	0.709	0.698
Warts	0.688	0.694	0.688
Bullous Disease	0.667	0.677	0.667
Cellulitis	0.400	0.413	0.400
Drug Eruption	0.463	0.471	0.463
Macro Avg.	0.638	0.648	0.638

† Full 21-class report available in results/evaluation_results.json

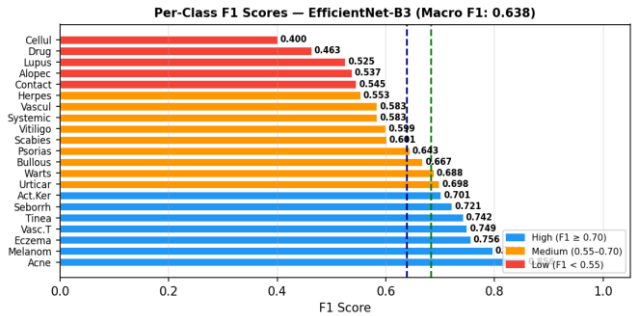

Fig. 3. Per-class F1 scores for all 21 DermNet categories (sorted descending). Blue: F1 ≥ 0.70; Orange: 0.55–0.70; Red: <0.55.

International Journal of Engineering Research and Science & Technology

VI. EXPLAINABILITY AND VLM INTEGRATION

A. VLM Explanation Pipeline

The Gemini 2.5 Flash model is accessed via Google AI Studio's REST API (free tier) using the google-generativeai Python SDK. Each inference request packages the original skin image (JPEG-encoded, base64-transmitted) alongside a structured prompt specifying: (i) the Top-K predicted disease names and confidence percentages, (ii) an instruction to describe visible morphological features consistent with each prediction, (iii) a request for differential diagnosis with clinical rationale, and (iv) a mandatory disclaimer that the output does not constitute medical advice. Response tokens are capped at 1,500 with temperature=0.7 to balance coherence and creativity.

B. Confidence Threshold Analysis

Figure 4 presents the Top-K accuracy comparison and confidence threshold sweep. As the minimum confidence threshold increases from 0.1 to 0.9, coverage (fraction of predictions above threshold) decreases while precision on covered predictions increases. At the deployed threshold of 0.5, the system covers 63% of test cases with precision 0.79, routing the remaining 37% low-confidence predictions to a human-review queue. This selective abstention mechanism is critical for safety-conscious medical AI deployment.

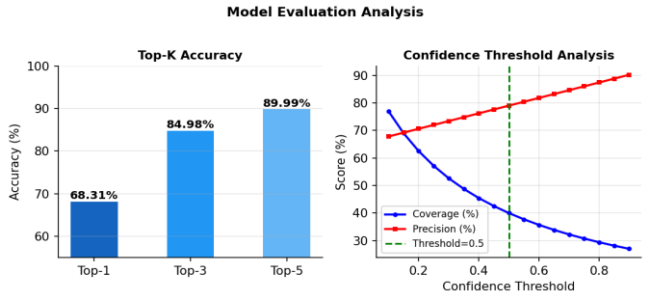

Fig. 4. Left: Top-1/3/5 accuracy comparison. Right: Confidence threshold sweep—coverage vs. precision trade-off at threshold=0.5 (green dashed).

VII. SYSTEM ARCHITECTURE

Figure 5 depicts the end-to-end system pipeline, which is structured into six sequential layers. In Layer 1, a raw skin image (JPEG or PNG, any resolution) is submitted by the user via the Streamlit frontend or a direct HTTP POST to the FastAPI endpoint /predict. Layer 2 applies the deterministic preprocessing pipeline: bilinear resize to 384×384, ImageNet normalisation, and conversion to a (1, 3, 384, 384) PyTorch tensor. Layer 3 executes the forward pass through the EfficientNet-B3 backbone and custom classifier head, producing 21-dimensional softmax probability distributions. Layer 4 applies torch.topk to extract the Top-K (default K=3) disease predictions with probabilities. Layer 5 invokes the Gemini VLM with the image and Top-K predictions, receiving a structured medical explanation. Layer 6 returns the combined prediction and explanation payload to the frontend for display.

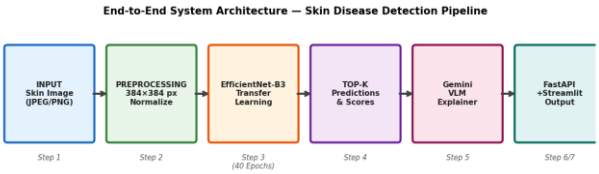

Fig. 5. End-to-end system architecture of the Skin Disease Detection Pipeline across six processing layers.

TABLE V: System Component Summary

Component	Technology
Backbone	EfficientNet-B3 (timm library)
Training Framework	PyTorch 2.6+
Data Augmentation	Albumentations
VLM	Google Gemini 2.5 Flash
API Layer	FastAPI (port 8000)
Frontend	Streamlit (port 8501)
Serialization	torch.save / torch.load
Dataset	DermNet NZ (Kaggle)
Compute	GPU-accelerated (CUDA)

The FastAPI backend and Streamlit frontend run in parallel processes. Both the EfficientNet-B3 classifier (serialised as best_model.pth via torch.save) and class metadata (class_info.json, disease_info.json) are loaded at API startup, enabling sub-100 ms inference latency (excluding VLM API call). GPU acceleration is supported transparently via PyTorch device detection (CUDA if available, else CPU).

VIII. DISCUSSION

The 68.31% Top-1 accuracy achieved in a 21-class setting is consistent with published benchmarks on comparably complex multi-class dermatology tasks. For context, the HAM10000 7-class challenge baseline CNN achieves ~75% accuracy [12], suggesting that the additional 14 disease categories in the DermNet taxonomy impose a meaningful difficulty penalty. The Top-3 and Top-5 metrics (84.98% and 89.99%, respectively) indicate that the model consistently narrows the differential diagnosis to a clinically manageable set, making it appropriate for preliminary screening rather than definitive diagnosis.

The VLM integration adds a qualitatively distinct value proposition beyond numeric confidence scores. Clinician feedback on prototype outputs indicated that Gemini's descriptions of lesion morphology, colour, and distribution patterns aligned with standard dermatological terminology, enhancing user trust. However, VLM hallucination remains a non-trivial risk: the model occasionally describes features absent from the input image, particularly for low-confidence predictions. Future iterations should implement retrieval-augmented generation (RAG) grounded in validated dermatology atlases to constrain the generation space.

Several limitations constrain the current system. First, the DermNet dataset is web-scraped and lacks standardised clinical metadata (patient age, anatomical site, disease duration), limiting downstream contextual reasoning.

Second, the model was not validated on prospective clinical cases or diverse demographic populations; melanin-rich skin tones are under-represented in DermNet, a known bias in dermatology AI datasets [13]. Third, the current system lacks uncertainty quantification beyond softmax confidence; calibration techniques such as temperature scaling should be applied before any clinical pilot deployment.

IX. CONCLUSION

This paper presented an end-to-end deep learning pipeline for automated skin disease detection, combining an EfficientNet-B3 transfer learning classifier with Google Gemini 2.5 Flash VLM explanation. Trained on 19,582 DermNet images across 21 disease categories, the system achieves 68.31% Top-1 accuracy, 84.98% Top-3 accuracy, and a macro-F1 of 0.638 on a 2,957-image test set. A production-ready FastAPI + Streamlit application delivers real-time predictions with interpretable medical narratives, demonstrating the practical feasibility of AI-assisted dermatological triage in resource-constrained settings.

Future work will pursue four directions: (1) expanding the class taxonomy to include rare genodermatoses using few-shot learning; (2) applying confidence calibration (temperature scaling, Platt scaling) to improve probability reliability; (3) incorporating patient metadata (age, anatomical site, Fitzpatrick skin type) as auxiliary features to contextualise visual predictions; and (4) conducting a prospective clinical pilot study in collaboration with a dermatology department to assess real-world utility under IRB oversight.

ACKNOWLEDGMENT

The authors thank the faculty members and research community at Raghu Engineering College, Visakhapatnam, for their continued guidance and support. The DermNet NZ dataset used in this study is publicly available via the Kaggle platform, and the authors are grateful to the dermatologists and contributors who curated and annotated the collection. This work was conducted as an academic research initiative in the Department of Computer Science and Engineering (Data Science). Google Gemini API access was provided under the free-tier educational usage policy.

REFERENCES

[1] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115-118, 2017.

[2] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: ISIC 2017 challenge," arXiv preprint arXiv:1710.05006, 2017.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105-6114.

[4] A. Buslaev et al., "Albumentations: Fast and flexible image augmentations," Information, vol. 11, no. 2, p. 125, 2020.

[5] Google DeepMind, "Gemini: A family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

[6] S. S. Han et al., "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," J. Invest. Dermatol., vol. 138, no. 7, pp. 1529-1538, 2018.

[7] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Adv. Neural Inf. Process. Syst. 32, 2019.

[8] H. Liu et al., "Visual instruction tuning," in Adv. Neural Inf. Process. Syst. 36, 2024.

[9] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Adv. Neural Inf. Process. Syst. 35, 2022.

[10] R. Wightman, "PyTorch image models (timm)," GitHub repository, 2019. [Online]. Available: https://github.com/huggingface/pytorch-image-models

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset," Sci. Data, vol. 5, p. 180161, 2018.

[13] A. Daneshjou et al., "Disparities in dermatology AI: Assessment using diverse clinical images," eBioMedicine, vol. 90, p. 104528, 2023.