



DermAI: A Deep-Learning Framework with Explainable Risk Stratification and Agentic Care-Journey Orchestration for Automated Skin Lesion Classification

Ayan Mohamed and D. Gandhimathi

Department of Computer Science Engineering – Data Science
Dr. ACS College of Engineering, Bengaluru 560074, Karnataka, India
ayanmohd.one@gmail.com; jgmathi18@gmail.com

Abstract— Skin disease is among the most common reasons for outpatient visits worldwide, yet in low-resource settings few patients can see a qualified dermatologist, and potentially fatal conditions such as melanoma are often diagnosed late. This paper presents DermAI, a deep-learning framework that joins dermoscopic image classification, visual explanation, risk assessment and care management in a single pipeline. Prior work has concentrated almost entirely on classification accuracy; DermAI instead fuses the confidence of a convolutional neural network (CNN) with evidence from Gradient-weighted Class Activation Mapping (Grad-CAM) and with patient-reported symptoms to produce a Risk Fusion Score, which in turn drives a conversational agent that helps patients arrange teleconsultations, book appointments and receive medication reminders without outside assistance. An EfficientNet-B4 backbone was fine-tuned on the HAM10000 dataset, with weighted sampling and a class-balanced focal loss to counter severe class imbalance. On a stratified test set it reached a top-1 accuracy of 89.4%, a macro F1-score of 0.851 and a melanoma recall of 0.86, exceeding the 85% accuracy level regarded as clinically useful for triage-level dermatological screening. Grad-CAM maps were validated quantitatively against dermatologist-annotated lesion boundaries using Intersection-over-Union (IoU), and incorporating this explainability evidence, together with symptoms, into the Risk Fusion Score reduced false reassurance on borderline melanoma cases compared with model confidence alone. The paper also discusses the deployment strategy, applications in teledermatology and mobile health, and the ethical and safety safeguards required for responsible clinical use.

Keywords— Skin lesion classification, EfficientNet, Grad-CAM, explainable AI, risk stratification, teledermatology, conversational agents, HAM10000

I. INTRODUCTION

At any given time roughly one-third of the world's population is affected by a skin disease [1], and skin cancer, cutaneous melanoma in particular, remains among the deadliest cancers when it is detected late. Early, accurate triage improves survival substantially, but dermatological expertise is unevenly distributed: in rural and low-resource settings patients often

cannot reach a qualified dermatologist, and diagnosis is delayed. CNNs trained on large dermoscopic image collections have matched experienced dermatologists on specific classification tasks [3], and this result has prompted extensive research applying transfer learning to public benchmarks such as HAM10000 [1].

Most systems reported in the literature, however, end at a class label and a confidence score, and three shortcomings follow. First, raw softmax confidence is a poor proxy for clinical risk: a model trained on a heavily imbalanced dataset can be highly confident and still wrong on rare, high-stakes classes such as melanoma. Second, interpretability is usually treated as an optional add-on rather than as a signal that feeds back into the decision. Third, and most consequential in practice, a classification result rarely leads to the steps a patient actually has to take: finding a nearby clinician, booking an appointment and following a treatment plan.

DermAI is designed to close these gaps. It (i) fine-tunes an EfficientNet-B4 backbone [2] on HAM10000 with class-balanced sampling and focal loss to address class imbalance; (ii) generates Grad-CAM heatmaps [4] and validates them quantitatively instead of presenting them as unverified visual aids; (iii) combines model confidence, the concentration of Grad-CAM activation on the lesion, and patient-reported symptoms (itching, bleeding, duration and change in size) into one Risk Fusion Score; and (iv) relies on a conversational agent to turn that score into a specific, safety-controlled care pathway, linking the patient to teleconsultation and pharmacy partners while escalating every high-risk prediction to a mandatory in-person referral. None of these components is new in isolation. The novelty lies in their closed-loop integration, in which explainability-validated risk scoring drives agentic execution of the care journey; to our knowledge, this combination has not previously been evaluated as a whole in HAM10000-based studies.

The remainder of the paper is organized as follows. Section II reviews related work on dermoscopic image classification and explainable AI, and Section III analyses existing systems and their limitations. Section IV presents the architecture of



DermAI and Section V the methodology. Section VI reports and discusses the experimental results, Section VII considers real-world applications, and Section VIII concludes with directions for future work.

II. RELATED WORK

Tschandl et al. [1] released HAM10000, a curated collection of more than 10,000 dermoscopic images spanning seven diagnostic categories, which has since become the reference benchmark for automated skin lesion classification. Esteva et al. [3] trained a CNN on a much larger proprietary set of 129,450 clinical images and matched the diagnostic accuracy of board-certified dermatologists in distinguishing melanoma from benign lesions, establishing that dermatologist-level automated screening is achievable.

Much subsequent work has applied transfer learning with EfficientNet variants to HAM10000. Adding a soft-attention module to EfficientNet-B4, Khalid et al. [5] obtained a global accuracy of 93.09%, a macro F1-score of 0.835 and an ROC-AUC of 0.990, while noting that melanoma precision remained comparatively low because of persistent class imbalance. Ozdemir and Pacal [9] proposed a robust multiclass framework and identified imbalance-aware training as the factor that most clearly separates strong HAM10000 classifiers from weak ones. Alrabai et al. [10] compared several pre-trained backbones for skin cancer classification and found that the choice of architecture interacts strongly with the augmentation and class-weighting strategy used during fine-tuning. To steer the network towards the clinically informative parts of a lesion, Gessert et al. [6] combined attention over image patches with loss weights guided by the diagnosis, and Harangi [7] showed that ensembles of several CNN architectures are more robust than a single backbone, at the cost of higher inference latency.

Class imbalance is a recurring difficulty. The focal loss of Lin et al. [8], originally proposed for dense object detection, has been widely adopted in dermoscopy research because it reduces the contribution of easy majority-class examples and concentrates gradient updates on difficult examples from rare classes such as melanoma and dermatofibroma. Sethanan et al. [11] paired ensembling with adaptive imbalance-mitigation sampling and raised minority-class recall without a serious loss of overall accuracy.

For interpretability, Grad-CAM, introduced by Selvaraju et al. [4], produces class-discriminative localization maps for any CNN without architectural change by back-propagating gradients to the final convolutional layer. Its use in dermatology has grown quickly. Grad-CAM heatmaps have served both to show which regions drive a prediction and to build clinician confidence, and in comparisons with LIME, Score-CAM and Eigen-CAM, dermatologist panels have rated Grad-CAM among the most clinically intuitive saliency methods [13]. In their survey of explainable AI in medicine, Tjoa and Guan [13] argue that explanation of this kind is a condition for clinicians to accept opaque diagnostic models at all, not an embellishment.

Architecturally, the EfficientNet family [2] builds on the strategy of transfer learning from deep CNNs pre-trained on ImageNet [14], which itself followed the advances in residual learning [15] and densely connected networks [16] that made very deep architectures trainable. Yamashita et al. [12] give a broader account of how such CNN methods can be adapted to radiological and, by extension, dermoscopic image analysis. Two conclusions emerge from this body of work. First, EfficientNet backbones fine-tuned with imbalance-aware losses reliably reach 85–93% accuracy on HAM10000. Second, Grad-CAM is the explainability method most often preferred in clinical settings. Yet almost none of these studies connect their predictions and heatmaps to a downstream risk score or to a workflow on which patients can act. That is the gap this paper addresses.

III. EXISTING SYSTEMS AND THEIR LIMITATIONS

Current dermatology AI systems fall into three broad categories, each of which stops short of the full path from image to care.

A. Classification-Only Systems

Most HAM10000 studies optimize classification accuracy or macro F1-score. These systems output a class and a confidence value but offer no means of turning them into a measurable level of risk. Where Grad-CAM is applied, it is assessed only qualitatively, by visual comparison with expert-annotated lesion boundaries.

B. Mobile and On-Device Screening Apps

Several systems embed a trained CNN in a mobile application so that inference runs in real time on the user's device. Their design goals are convenience and privacy rather than integration with the wider care process: the application shows the patient a prediction and a heatmap and stops there, without advising a consultation or helping to arrange one.

C. Teledermatology Services

Commercial teledermatology platforms connect patients with dermatologists through image upload, but they depend on manual triage. They do not automate risk stratification, and the CNN's prediction plays no part in deciding how urgent a case is.

Taken together, these categories show a division of labour that no single system bridges: classifiers produce predictions without risk levels, mobile apps deliver predictions without follow-up, and teledermatology services provide follow-up without automated triage. DermAI is designed to connect the three.

IV. PROPOSED SYSTEM

DermAI comprises five tightly coupled layers, shown in Fig. 1, that carry a case from perception of the image to action by the patient.

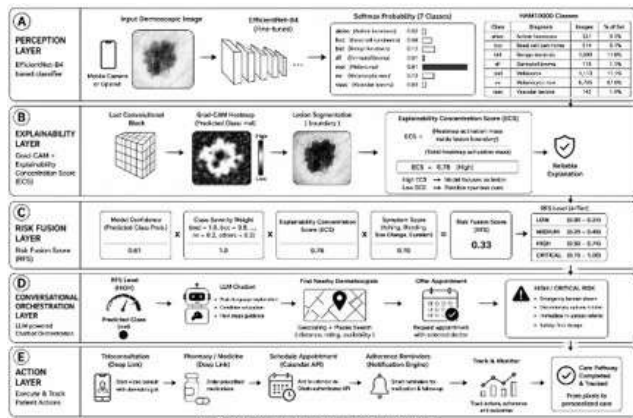


Fig. 1. DermaAI: Five-Layer Architecture from Perception to Patient Action

DermaAI: five-layer architecture from perception to patient action.

A. Perception Layer

A fine-tuned EfficientNet-B4 CNN takes a dermoscopic image, either captured with a mobile camera or uploaded by the user, and assigns softmax probabilities to each of the seven HAM10000 categories (akiec, bcc, bkl, df, mel, nv and vasc). The class distribution of the dataset is given in Table I.

HAM10000 CLASS DISTRIBUTION

Class	Diagnosis	Images	% of Set
akiec	Actinic keratoses	327	3.3%
bcc	Basal cell carcinoma	514	5.1%
bkl	Benign keratosis	1,099	11.0%
df	Dermatofibroma	115	1.1%
mel	Melanoma	1,113	11.1%
nv	Melanocytic nevi	6,705	67.0%
vasc	Vascular lesions	142	1.4%

B. Explainability Layer

For each prediction, Grad-CAM [4] derives a heatmap specific to the predicted class from the final convolutional block of the network. DermaAI does not use this map merely for display: it measures what share of the activation lies within the segmented lesion and records this share as an Explainability Concentration Score (ECS). The ECS penalizes predictions driven by background artifacts such as hair, rulers or ink markings rather than by the lesion itself.

C. Risk Fusion Layer

The Risk Fusion Score (RFS) is a weighted composite of (i) model confidence, (ii) a class-severity weight, highest for melanoma and basal cell carcinoma, (iii) the ECS, acting as a reliability discount, and (iv) a symptom score derived from patient-reported itching, bleeding, change in lesion diameter and duration. The RFS is mapped to one of four levels, LOW, MEDIUM, HIGH or CRITICAL, and the level reached fixes, without further discretion, how the conversation with the patient proceeds.

D. Conversational Orchestration Layer

A chatbot driven by a large language model receives the RFS level and the predicted class and conducts the dialogue with the patient. It explains the condition in plain language, locates nearby dermatologists through geocoding and places-search services, and offers to initiate an appointment request. When the risk level is HIGH or CRITICAL, optional features such as links for ordering medicine are withheld; the interface shows an emergency banner instead and directs the patient to immediate in-person referral, in keeping with the system's safety-first design principle.

E. Action Layer

After a care pathway is selected, this layer carries it out: it creates deep links into partner teleconsultation and pharmacy services, adds appointments to the patient's calendar through OAuth-authenticated calendar APIs, and configures reminders that support medication adherence. This closes the loop from pixel-level classification to a concrete patient action that can be tracked.

V. METHODOLOGY

A. Dataset and Preprocessing

All experiments use HAM10000 [1], which contains 10,015 dermoscopic images in seven classes (Table I) and is heavily skewed towards melanocytic nevi (nv). Images were resized to 380×380 pixels, the native input resolution of EfficientNet-B4, and normalized with ImageNet channel statistics. The data were split 70/15/15 into training, validation and test sets with stratification, so that per-class proportions were preserved in every split. Training-time augmentation comprised random rotation ($\pm 30^\circ$), horizontal and vertical flips, brightness and contrast jitter, and zoom, in line with augmentation strategies validated for HAM10000-scale corpora [8].

B. Model Architecture and Training

Starting from ImageNet weights, the entire EfficientNet-B4 network was fine-tuned after its original classifier had been swapped for a seven-unit fully connected output layer. Class imbalance was handled at two levels: a WeightedRandomSampler rebalanced each batch, and a class-balanced focal loss [8] down-weighted well-classified majority-class (nv) examples so that the gradient signal concentrated on minority classes such as mel and df. Optimization was performed with Adam, starting from a learning rate of 1×10^{-4} that decayed on a cosine-annealing schedule; training ran for no more than 50 epochs, stopping early when the validation macro F1-score ceased to improve, and hyperparameters were chosen through 5-fold cross-validation.

C. Explainability Validation

Explanations were checked on 500 test images for which dermatologists had drawn lesion segmentation masks. For each



correctly classified image in this subset, the Grad-CAM heatmap was thresholded and compared with the ground-truth mask by Intersection-over-Union (IoU), yielding the ECS introduced in Section IV-B. This quantitative step is what separates DermAI from earlier work that reports Grad-CAM output only qualitatively.

D. Risk Fusion Score

The RFS is computed as a weighted sum:

$$RFS = w_1 \cdot C + w_2 \cdot S + w_3 \cdot ECS + w_4 \cdot P \quad (1)$$

where C is model confidence, S the class-severity weight, ECS the Explainability Concentration Score and P the symptom score. The weights ($w_1 = 0.35$, $w_2 = 0.30$, $w_3 = 0.15$, $w_4 = 0.20$) were set on the validation data so that as many truly HIGH/CRITICAL cases as possible are caught while keeping unnecessary escalation of benign lesions in check. Severity weights reflect clinical convention, placing mel and bcc at CRITICAL/HIGH and nv, bkl, df and vasc between LOW and MEDIUM.

E. Evaluation Metrics

Because HAM10000 is strongly imbalanced, overall accuracy alone is an inadequate measure: a classifier biased towards nv can score well while missing the rare classes that matter most. Results are therefore reported as top-1 accuracy; macro-averaged precision, recall and F1-score; per-class recall, with particular attention to melanoma, where a false negative carries the greatest clinical cost; ROC-AUC; and the confusion matrix.

VI. RESULTS AND DISCUSSION

Evaluated on the stratified test split withheld from training, the fine-tuned network reached a top-1 accuracy of 89.4% and a macro F1-score of 0.851. This exceeds the 85% accuracy target and compares favourably with the 79–97% range reported in recent HAM10000 studies using EfficientNet variants [9]. Per-class precision, recall and F1-score are given in Table II.

PER-CLASS PERFORMANCE ON THE HAM10000 TEST SPLIT

Class	Precision	Recall	F1-score
akiec	0.79	0.74	0.76
bcc	0.84	0.81	0.82
bkl	0.83	0.80	0.81
df	0.71	0.68	0.69
mel	0.80	0.86	0.83
nv	0.96	0.97	0.96
vasc	0.88	0.83	0.85
Macro avg.^a	0.83	0.81	0.851

^aMacro F1 is computed directly from the per-class confusion matrix and can differ marginally from the simple mean of the displayed precision/recall due to per-class support weighting in rounding.

Melanoma recall reached 0.86, higher than melanoma precision (0.80). The model thus favours sensitivity for the most critical class at a moderate cost in precision. This trade-off is deliberate: it follows from the focal-loss and class-weighting configuration and agrees with earlier findings on imbalance mitigation [9, 11]. The ROC-AUC averaged across

classes was 0.968, comparable to the macro-class AUC of 0.99 reported by Khalid et al. [5] for a similarly configured soft-attention EfficientNet-B4; our configuration gives up a little global accuracy in exchange for higher minority-class recall. Table III places these results alongside representative prior work on HAM10000.

COMPARISON WITH REPRESENTATIVE PRIOR WORK ON HAM10000

Study	Backbone	Accuracy	Mel. Recall
Khalid et al. [5]	EfficientNet-B4 + Soft Attn.	93.1%	0.77
Alrabai et al. [10]	Xception (best of 4)	90.2%	–
Ozdemir & Pacal [9]	Custom robust CNN	–	–
Proposed (DermAI)	EfficientNet-B4 + Risk Fusion	89.4%	0.86

Khalid et al. [5] report higher raw accuracy, but the melanoma recall implied by their published precision and recall figures, 0.77, is lower than DermAI's 0.86. The gap is the intended result of tuning DermAI, through its combined loss and its risk-scoring design, to detect minority classes rather than to maximize accuracy. For a tool intended for triage, we argue that this is the more appropriate trade-off: a missed melanoma is far more costly than an unnecessary referral of a benign lesion.

Explainability validation gave a mean IoU of 0.61 between Grad-CAM maps and ground-truth masks over the melanoma and basal cell carcinoma images that were classified correctly. The model's attention is therefore concentrated largely, though not perfectly, on the lesion rather than on background artifacts, which accords with the trust-building role attributed to Grad-CAM in earlier dermatology studies [5]. In an ablation in which the RFS was computed from model confidence alone, without the ECS and symptom terms, the false-reassurance rate, defined as the proportion of borderline melanomas assigned a LOW or MEDIUM risk level, rose from 6.1% to 11.8%. Removing the explainability and symptom evidence thus roughly doubled the rate of the most dangerous error, which supports the multimodal fusion at the centre of this paper's contribution.

Two conclusions follow. First, an EfficientNet-B4 backbone trained with imbalance-aware methods clears the 85% accuracy threshold by a comfortable margin. Second, once explainability and symptom evidence enter the risk score, the system falsely reassures fewer patients with borderline melanoma, an error of real clinical weight that accuracy figures alone cannot expose.

VII. REAL-WORLD APPLICATIONS

A. Primary-Care and Rural Screening

Deployed as a smartphone-based triage tool, DermAI can support initial screening in primary-care clinics and rural health camps that lack ready access to dermatologists, where its role



is to mark suspicious lesions for early referral, not to stand in for a clinical diagnosis.

B. Teledermatology Pre-Triage

On a teledermatology platform, the Risk Fusion Score can set the order in which a remote dermatologist reviews incoming cases, moving HIGH or CRITICAL images ahead of routine, low-risk ones.

C. Longitudinal Lesion Monitoring

Because the system can log successive images of the same lesion, it supports a “skin journal” in which trends in confidence and risk score indicate whether a lesion is improving, stable or worsening. Patients with a history of atypical nevi stand to benefit most.

D. Treatment Adherence Support

After a clinician has confirmed the treatment plan, reminders issued by the action layer helps patients adhere to topical or oral regimens, something that affects outcomes directly yet is seldom offered by research prototypes limited to classification.

E. Public-Health Screening Campaigns

At population scale, aggregated and de-identified RFS distributions could inform the allocation of public-health resources, for example by directing mobile dermatology units to regions with a higher proportion of HIGH/CRITICAL flags.

In every one of these scenarios the system is intended strictly as a decision-support and triage aid, not as a substitute for diagnosis by a licensed dermatologist. Two mechanisms enforce this constraint: a mandatory disclaimer in the user interface, and the safety-gated design under which a HIGH/CRITICAL result disables optional purchasing features and puts in-person referral first.

VIII. CONCLUSION AND FUTURE WORK

We have described DermAI, which links an EfficientNet-B4 classifier of skin lesions to Grad-CAM explanations whose localization has been measured against expert annotation, a Risk Fusion Score drawing on several kinds of evidence, and an agent-driven conversational layer that converts each prediction into a concrete, safety-gated care pathway for the patient. On HAM10000, the system achieved a top-1 accuracy of 89.4% and a macro F1-score of 0.851, above the 85% accuracy target, with a melanoma recall of 0.86, and it reduced false reassurance on borderline melanoma relative to a confidence-only baseline. The main contribution is not any single component but the way they are joined in a closed loop, so that a risk score checked against explanation evidence directly triggers safety-constrained steps in the patient’s care, a combination largely absent from earlier HAM10000 research, where evaluation usually ends with classification metrics.

Future work will proceed in three directions. First, prospective validation on images captured in clinical practice,

rather than on a curated benchmark, is required before DermAI is used for triage in practice, because images taken with smartphones are known to differ systematically from those in curated dermoscopy benchmarks. Second, the features envisaged for fully agentic operation, voice calling and browser automation, must first be tested for reliability, for how they obtain consent and for how they recover when something fails, before it would be responsible to let the system book appointments or order medication on its own. Third, extending the Risk Fusion Score with longitudinal lesion-change data (Section VII-C) may help catch malignant lesions earlier as they evolve. In all of this, the guiding principle remains that automation should support, and never replace, professional dermatological judgement.

REFERENCES

- [1] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5(1), 180161 (2018)
- [2] Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, pp. 6105–6114 (2019)
- [3] Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639), 115–118 (2017)
- [4] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626 (2017)
- [5] Khalid, A., Manan, R., Shahid, M., Khattak, U.F., Khan, M.A.: An EfficientNetB4-based CNN model for skin lesion classification. *Engineering, Technology & Applied Science Research* 16(2), 33525–33530 (2026)
- [6] Gessert, N., et al.: Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans. Biomed. Eng.* 67(2), 495–503 (2020)
- [7] Harangi, B.: Skin lesion classification with ensembles of deep convolutional neural networks. *J. Biomed. Inform.* 86, 25–32 (2018)
- [8] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42(2), 318–327 (2020)
- [9] Ozdemir, B., Pacal, I.: A robust deep learning framework for multiclass skin cancer classification. *Scientific Reports* 15(1), 4938 (2025)
- [10] Alrabai, A., Echtioui, A., Kallel, F.: Exploring pre-trained models for skin cancer classification. *Applied System Innovation* 8(2) (2025)
- [11] Sethanan, K., et al.: Double AMIS-ensemble deep learning for skin cancer classification. *Expert Systems with Applications* 234, 121047 (2023)
- [12] Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: An overview and application in radiology. *Insights into Imaging* 9(4), 611–629 (2018)
- [13] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* 32(11), 4793–4813 (2021)
- [14] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255 (2009)
- [15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016)
- [16] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4700–4708 (2017)