JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DERMAAI: A CALIBRATED DEEP-LEARNING WEB FRAMEWORK FOR SKIN CONDITION SCREENING AND ATOPIC DERMATITIS SEVERITY ESTIMATION

Transfer learning with EfficientNet-B0, temperature-scaled confidence and a service-oriented web design

[1]Zainab Fathima, [2]Seema Kumari, [3]Sahana Pujari, [4]Srujan AB, [5]Sunil A

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor
[1-5]Department of Computer Science and Engineering,
[1-5]City Engineering College, Bengaluru, India

Abstract : A photograph of a rash rarely settles whether the cause is atopic dermatitis (AD), and a class label on its own tells the reader nothing about how much to trust it. DermaAI is a web-based screening framework that pairs a fine-tuned EfficientNet-B0 image classifier with explicit treatment of uncertainty and a transparent severity grade. After a single photograph is uploaded, the system reports the leading condition, the three most likely classes, a temperature-calibrated confidence, a low-confidence warning and a mild, moderate or severe visual grade. A React front end and an Express back end are kept apart from a FastAPI inference service, so a retrained model can be swapped in without changing the application. On 138 held-out images from six classes the classifier achieved 71.74% accuracy and a macro F1 of 70.23%. The fitted calibration temperature was 1.030532, and any prediction below a confidence of 0.55 is flagged for professional review. A complete run on a real photograph returned Atopic Dermatitis with 82.52% confidence. At present the severity grade is derived from the share of visibly red skin rather than from a trained model, and a designed image-plus-clinical fusion network has not yet been validated; both limits are stated plainly. The findings show that a compact network can support a careful, traceable screening workflow, although it is not yet reliable enough for clinical use.

IndexTerms - atopic dermatitis, skin disease classification, EfficientNet, transfer learning, confidence calibration, severity estimation, teledermatology

I. INTRODUCTION

Atopic dermatitis, commonly called eczema, is a long-lasting inflammatory skin disease marked by dry, itchy and reddened patches that appear, fade and return over many years. Managing it well depends on two judgements: whether a given lesion really is AD, and how severe the current flare is. Both are still made largely by visual inspection. Two specialists can reach different conclusions about the same patch, and in many regions a dermatologist is far away.

The effects are not only physical. Visible skin disease is associated with anxiety, low mood and withdrawal from social life, and global burden studies rank skin conditions among the main sources of non-fatal health loss [1]. Early treatment usually works well, so a missed or mistaken diagnosis carries a real cost. The difficulty is that several disorders look alike: eczema, psoriasis and contact dermatitis share colour, scaling and body site, which blurs the boundaries even for experts.

Deep learning has changed how skin images are analysed. Convolutional networks trained on large dermatology collections have equalled specialists on narrow tasks such as lesion triage [2]. Yet many published systems return only a label. A person using a screening tool also needs to know how sure the model is, whether the photograph was usable and what the tool cannot do. Studies of dermatology models report weaker results on darker skin and on rare image types [3], which makes honest communication of uncertainty a matter of safety.

A. Problem Statement

Three gaps motivate this work. First, most classifiers rely on a photograph alone, whereas a clinician would also ask about history, itch and duration. Second, the raw softmax scores of modern networks tend to be overconfident [4], so a high score does not mean a high chance of being right. Third, established severity indices such as EASI [5] and SCORAD [6] need a trained observer, which restricts their use in remote or high-volume settings.

B. Contributions

- An end-to-end screening pipeline in which one browser upload yields a condition, a top-three list, a calibrated confidence and a severity grade in a single request.
- A service-oriented design that places the model behind an HTTP interface, so it can be retrained, versioned or moved to faster hardware independently of the user-facing application.
- A cautious presentation layer combining temperature scaling, a 0.55 low-confidence threshold, labelling of severity as an automated visual estimate, and a model-version record stored with every result.
- An open statement of what has and has not been validated: the deployed severity grade is a colour-based heuristic, and the image-plus-clinical fusion model is a design that is excluded from the reported accuracy.

The remainder of the paper is organised as follows. Section II surveys related work, Section III presents the architecture and methods, Section IV describes the implementation, Section V reports the experiments, Section VI discusses the findings, Section VII covers limitations and ethics, and Section VIII concludes.

II. Related Work

A. Image-Based Skin Disease Classification

A deep network trained on about 130,000 clinical images reached dermatologist-level performance in separating malignant from benign lesions [2], and public dermoscopy collections such as HAM10000 [11] allowed many groups to train and compare models. Most of this research addresses lesion triage, for example distinguishing melanoma from a mole, and depends on large, expertly labelled datasets. Inflammatory diseases are harder because their appearance varies with body site, stage and skin tone, and well-labelled images are scarcer.

B. Automated Severity Scoring for Atopic Dermatitis

In clinical practice AD severity is expressed with the Eczema Area and Severity Index [5] and the SCORAD index [6], which combine how much skin is involved with the intensity of signs such as redness and scaling. One team trained an ensemble of convolutional networks on several thousand labelled photographs to estimate EASI and reported agreement comparable with clinicians [7]. A pilot study applied deep learning to SCORAD-style grading and found a promising correlation with clinical scores, although its dataset was small [8]. Such systems target the clinical index directly, but they require large expert-labelled collections that a student or community project seldom has. For this reason DermaAI offers a simpler visual estimate and states that it is not a clinical score.

C. Multimodal Learning

A photograph records what the skin looks like but not what the patient has experienced. Multimodal learning seeks to merge such complementary sources [10]. In dermatology, combining an image encoder with metadata such as age, sex and lesion site has improved classification over images alone [9], and comparisons of fusion architectures for skin lesions continue to appear [20]. Recent AD-specific work describes multimodal methods that join image and clinical inputs [13], and vision-language models have been explored for automated AD grading [14]. These studies suggest that clinical context can separate look-alike conditions, but they also caution that fusion adds parameters and can overfit when data are limited. Our design therefore treats simple concatenation as the baseline that any richer fusion must outperform.

D. Fairness and Reliability

Tests on images labelled by Fitzpatrick skin type show uneven accuracy across tones [12], and a curated clinical image set exposed weaker results on darker skin and on rarer diseases [3]. Independently, modern networks are often poorly calibrated, and a simple post-hoc correction known as temperature scaling brings confidence closer to actual accuracy [4]. These findings explain our focus on calibration, a conservative abstention threshold and per-group testing as a condition for wider release.

table 1. positioning of dermaai against related approaches

Aspect	Typical lesion classifiers	AD severity scorers	DermaAI (this work)
Output	Class label	EASI or SCORAD estimate	Class, top-3, calibrated confidence, severity grade
Data need	Large labelled sets	Large expert-scored sets	About 150 images per class
Uncertainty shown	Rarely	Rarely	Yes, with a 0.55 flag
Severity basis	None	Clinical index	Visual redness share, labelled non-clinical
Deployment	Research code	Research code	Web application with separate inference service

III. System Architecture and Methodology

A. Overview

DermaAI consists of three cooperating tiers, kept separate on purpose so that the component that changes most often, the trained model, sits behind a narrow interface. The presentation tier is a React single-page application in which a user registers, signs in, uploads or captures a photograph, completes a short clinical form and reads the result. The application tier is a Node.js and Express back end that authenticates users, validates and stores uploads, forwards the image to the inference service, derives the severity grade and saves each prediction in MongoDB. The inference tier is a Python FastAPI service that loads the trained PyTorch checkpoint, preprocesses the image, runs the network, calibrates the scores and returns JSON.

The tiers communicate through JSON over HTTP. Every response includes the model version, the backbone name and the calibration temperature, so any stored prediction can be traced to the exact model that produced it.

B. Image Classifier

The classifier is EfficientNet-B0 [15], a compact convolutional network whose depth, width and input resolution are scaled together. We initialise it with ImageNet weights [16] and adapt it by transfer learning, which reuses general edge, texture and colour detectors and leaves the skin-specific details to be learned from our data. The original classification layer is removed, and the pooled feature vector f passes through a small head given by Eq. 1 and Eq. 2.

$$z = W_2 \, \mathrm{Drop}_{0.4}(\mathrm{ReLU}(W_1 \, \mathrm{Drop}_{0.3}(f))) \quad (1)$$

$$\hat{p} = \mathrm{softmax}(z / T) \quad (2)$$

Here W_1 maps the features to 512 units, W_2 maps to the class scores z, and T is the calibration temperature introduced in Section III-D. The number of outputs is read from the training folders, so adding a condition needs new labelled images and retraining but no code change. The backbone is configurable and can be replaced by a larger EfficientNet or a ResNet.

C. Data Preparation and Training

Images are arranged by class. A preparation script removes duplicates and unreadable files, discards poor-quality pictures and selects a balanced subset for each class. The data are divided 70% for training, 15% for validation and 15% for testing using a fixed seed of 42, and the test split is never used to fit weights or choose a checkpoint. A check confirms that all three splits contain the same classes and warns when class counts differ greatly.

Training augmentation resizes and randomly crops each image, flips and rotates it, jitters colour, applies small shifts, normalises with ImageNet statistics and erases random patches, imitating the variety of real phone photographs in lighting, angle, framing and occlusion. At evaluation and serving time only a deterministic resize, centre crop and normalisation are used, and the same preprocessing code serves both training and inference so that the model never meets a different input distribution in production. Training minimises cross-entropy with the Adam optimiser [18], uses early stopping on validation performance, and follows two phases: the new head is trained first with the backbone frozen, after which the backbone is unfrozen and fine-tuned at a lower learning rate. The checkpoint with the best validation score is kept.

D. Confidence Calibration and Abstention

Softmax outputs from deep networks are frequently too confident [4]. We therefore divide the logits by one scalar T before the softmax, fitting T on the validation set by minimising negative log-likelihood. A value of T above 1 softens the distribution and a value below 1 sharpens it. Since only one parameter is learned, the class ranking and the accuracy are unchanged; only the meaning of the probability shifts.

The calibrated probability of the top class is then compared with a threshold of 0.55. When it falls below, the result is marked as low confidence and the interface advises consulting a dermatologist. The threshold is configurable and should eventually be set from validation data and per-group reliability curves.

E. Severity Estimation

Two severity routes exist and are kept distinct. The image-based visual estimate, which is deployed, measures the share r of visible skin whose colour indicates elevated redness and compares r with cut-points, so low values map to mild, intermediate values to moderate and high values to severe. In our logged test r was about 26.3%, which the system graded as high, and the response states that this is an automated visual assumption rather than a clinical score. The clinical-form heuristic, also deployed in the back end, derives a grade from questionnaire fields such as symptoms and duration through fixed rules rather than a learned head. Neither route replaces EASI or SCORAD, and the interface presents both as decision-support cues.

F. Designed Multimodal Fusion Model

Beyond the deployed path we designed and prototyped a joint model with two encoders and two output heads. A convolutional encoder produces an image vector, and a small dense encoder converts clinical fields (age, itch level, affected area, history) into a metadata vector after range checks, scaling with training-set statistics and explicit handling of missing values. The two vectors are concatenated and passed through a shared layer to a diagnosis head and a three-class severity head. Training minimises the weighted sum in Eq. 3, with the weights chosen on validation data.

$$L = \lambda_d \, L_{diag} + \lambda_s \, L_{sev} \quad (3)$$

A prototype was written with a MobileNetV2 image branch. It has not been trained on enough paired image and clinical data, so no accuracy is claimed for it. Before any benefit is asserted, a fusion model should be compared with an image-only baseline using identical patient-level splits.

IV. Implementation

A. Technology Stack

Table 2 lists the components. The model code uses PyTorch [17], and the other choices favour widely supported tools that are easy to deploy.

table 2. software stack

Layer	Technology	Responsibility
Frontend	React 18, React Router 6, Vite 5, Axios	Upload, camera capture, forms, results, history, dashboard
Backend	Node.js, Express 4, Mongoose 8, Multer	Routing, file handling, severity rules, persistence
Security	JWT, bcrypt, Helmet, rate limiting, input sanitisation, CORS	Authentication and request hardening
Database	MongoDB Atlas	Users, predictions, disease pages, chat records
Inference	Python, FastAPI, Uvicorn	Preprocessing, prediction, calibration
Model	PyTorch, EfficientNet-B0	Image classification
Assistant	Hosted large language model	Educational chatbot on general skin health

B. Request Flow

One analysis passes through eight steps. (1) The user selects or captures an image and submits the clinical form. (2) The back end verifies the session token and checks the file type and size, rejecting invalid or empty files. (3) The image is saved under a unique name. (4) The back end posts the image to the inference service. (5) The service resizes the image to 224 by 224 pixels, normalises it, runs the network and divides the logits by the fitted temperature. (6) The service returns the condition, calibrated confidence, top three classes, low-confidence flag, severity text with its reason, and model details. (7) The back end merges this with the clinical-form heuristic, stores the record and returns it. (8) The interface displays the result beside a statement that it is not a medical diagnosis. If the inference service cannot be reached, the back end returns a plain error and shows no prediction, so a failure is never mistaken for a result.

C. Response Contract

The inference service answers with a JSON object whose interpretation and audit fields are listed in Table 3.

table 3. fields of the prediction response

Field	Meaning
condition	Most likely class
confidence	Calibrated probability between 0 and 1
is_low_confidence	True when confidence is below the threshold
confidence_threshold	Threshold applied (0.55)
severity, severity_reason	Visual severity grade and its plain-language basis
model_version, backbone	Identity of the model that produced the answer
temperature	Calibration temperature applied
demo_mode	False when a real model produced the result

The demonstration flag matters because the application can run in a demo mode that returns clearly labelled placeholder predictions, allowing the interface to be built before a model exists. Every result reported in this paper was produced with that mode switched off.

D. Interface Safeguards and Additional Modules

The results page is designed to discourage over-reading. It presents confidence as a model estimate rather than certainty, repeats the disclaimer beside the condition, labels severity as automated and displays the low-confidence warning prominently. Passwords are stored as bcrypt hashes and sessions use signed tokens. Uploads are validated by type and size, logs record technical diagnostics without patient identifiers, and the inference service refuses inputs and configuration files that do not match the schema saved at training time. The application also offers a per-user prediction history, informational pages for supported conditions, a dashboard of total scans, the most common condition and the count of low-confidence results, and an educational chatbot. These modules aid usability but do not influence the reported classification results.

V. Experiments and Results

A. Experimental Setup

The classifier was trained and tested on a Windows workstation using Python and PyTorch. The dataset contains about 150 images per skin-condition class, collected from public dermatology collections and cleaned as described in Section III-C. After the 70/15/15 split, 138 images were held out for testing, and these played no part in fitting weights, selecting the checkpoint or fitting the temperature. The metrics were regenerated from the saved checkpoint by an evaluation script, so the figures below are measured rather than estimated. We report accuracy and macro-averaged F1, which gives every class equal weight so that a large class cannot hide weak performance on a small one.

B. Classification Performance

table 4. held-out test performance of the six-class model

Metric	Value	Test images
Accuracy	71.74%	138
Macro F1	70.23%	138
Number of classes	6	–
Backbone	EfficientNet-B0	–

The model labelled roughly seven of every ten unseen images correctly. Macro F1 lies close to accuracy, which suggests that no single dominant class drives the result. For about 150 images per class we regard this as encouraging for a screening aid, yet clearly short of what independent clinical use would demand. With 138 test images, one image shifts accuracy by about 0.7 percentage points, so the estimate is uncertain and should be re-measured on a larger external set.

C. Calibration and Abstention

Temperature scaling on the validation set gave T = 1.030532. A value so near 1 indicates that the raw scores were already close to calibrated, so the correction is small and the reported confidences are not inflated. Predictions whose calibrated confidence is below 0.55 are flagged. A reliability diagram and the expected calibration error on the test split have not yet been produced and are listed as pending work.

D. End-to-End Pipeline Test

To verify the whole system, a real photograph was uploaded through the web application with demonstration mode off, and each stage was inspected. The outcome is summarised in Table 5.

table 5. result of the end-to-end test

Check	Expected	Observed	Result
Image stored by back end	File saved in uploads folder	Stored as a JPEG	Pass
Inference service called	Request reaches prediction endpoint	Request received	Pass
Prediction returned	Condition and confidence	Atopic Dermatitis, 82.52%	Pass
Real model used	demo_mode false	demo_mode false	Pass
Threshold logic	0.8252 above 0.55, so not low confidence	is_low_confidence false	Pass
Severity returned	Grade with reason	High, about 26.3% red area	Pass
Traceability	Version and backbone present	v4.0.0, efficientnet_b0	Pass

Every field was present and consistent. A confidence of 0.8252 cleared the 0.55 threshold comfortably, so the result appeared as a regular prediction, and the severity text explained that it rested on the proportion of visibly red skin. This single run shows that the pipeline works end to end; it is evidence of integration and not of accuracy, which Table 4 addresses.

E. Software Testing

The system was also exercised at several levels. Unit tests cover resizing to 224 by 224 pixels, pixel normalisation, label mapping and confidence formatting. Integration tests confirm that an uploaded image reaches the model and that the stored record matches the response. Functional tests check that JPEG and PNG files are accepted, that PDF and executable files are rejected, that duplicate registrations and wrong passwords are refused, and that the correct condition and confidence appear on screen. Security checks verify that passwords are hashed, that a file renamed to look like an image is rejected, and that one user cannot open another user's report by editing the address. Regression tests are repeated after each retraining or interface change.

VI. Discussion

The main finding is that a lightweight network, fine-tuned on a modest dataset, can drive a complete screening workflow that is fast enough for interactive use and careful in how it addresses the user. An accuracy of 71.74% is not high in absolute terms, but it was obtained with about 150 images per class and little tuning. The more transferable result is the surrounding design: the same classifier shown as a bare label would invite over-trust, whereas calibrated confidence, an abstention flag and an explicit disclaimer give both user and clinician what they need to weigh the answer.

Calibration and abstention matter because people who use a screening tool cannot check its output. If a network reports 95% when it is right only 70% of the time, users will act on advice that deserves doubt. The fitted temperature of 1.030532 shows that the problem was mild for this model, which is itself worth reporting. Even so, calibration on a small validation set can be unstable, and one global threshold treats all classes and skin tones alike. A sturdier approach would add per-class reliability analysis and an out-of-distribution check so that a landscape or non-skin image is refused instead of labelled.

The redness-share grade needs no extra labels and is easy to explain, but it has clear weaknesses. Redness is harder to detect on darker skin, white balance and lighting change the measured share, and a red area is not the same as a clinically scored lesion. The grade can therefore vary for reasons unrelated to the disease, and we treat it as a coarse prompt to seek care. Validation against dermatologist scores is the most important open item.

Although the project is motivated by multimodal learning, the reported accuracy comes from an image-only network, and clinical-form data enter only through the severity heuristic. We prefer to say so than to imply that fusion was demonstrated. The motivation remains sound, since history, itch and duration are exactly what a clinician uses to separate look-alike conditions, and earlier studies report gains from combining them with images [9], [13]. A fair test needs paired image and clinical data, identical patient-level splits and comparison with the image-only baseline reported here.

Separating the inference service from the web application also paid off in practice. The model could be retrained and replaced without touching the interface, the version recorded in each response allowed outputs to be compared across retraining rounds, and sharing preprocessing between training and serving removed a common source of silent error.

VII. Limitations and Ethical Considerations

A. Limitations

- Small, single-source evaluation: the test set has 138 images and no external site, so accuracy may not carry over to new cameras, clinics or populations.
- Narrow label space: the model knows only the classes it was trained on and will still output a label for a condition outside that set. Atopic Dermatitis has since been added as a target class, and the model must be retrained on the enlarged dataset before the reported figures apply to it.
- Unmeasured subgroup performance: results have not been compared across skin tones, ages, body sites or lighting, although other dermatology models suggest such gaps are likely [3], [12].
- Severity is not clinical: the redness-share grade and the questionnaire heuristic have not been compared with EASI or SCORAD scored by dermatologists.
- Calibration not fully assessed: a reliability diagram and expected calibration error on the test split remain to be produced.
- Limited image-quality control: blurred, dark, distant or filtered photographs can degrade predictions, and unsuitable images are not yet reliably rejected before prediction.
- Local deployment and no fusion result: testing ran on one machine, so behaviour under many simultaneous users is unknown, and multimodal fusion is designed but not evaluated.

B. Ethical and Regulatory Considerations

Skin photographs are sensitive health data. Any deployment should encrypt images in storage and in transit, restrict access by role, define retention periods and let users delete their data. Using images to improve the model should be opt-in and explained in plain language. Explainability tools such as Grad-CAM [19] can help developers notice shortcuts, for example a model keying on a ruler or background, although they do not prove that the model reasons in a medically valid way.

DermaAI is a screening and awareness aid. It is not a medical device, it does not diagnose, and it must not be used to start or stop treatment. If offered to the public as a diagnostic aid it could fall under medical-device and health-data regulation, including the Central Drugs Standard Control Organisation rules and the Digital Personal Data Protection Act in India, which should be reviewed with a legal or regulatory expert before release. Uncertain and serious cases should be routed to a qualified dermatologist rather than decided by the model alone.

VIII. Conclusion and Future Work

This paper presented DermaAI, a web-based framework that converts one skin photograph into a predicted condition, a top-three list, a calibrated confidence, a low-confidence warning and a visual severity grade, with the model version recorded in every response. An EfficientNet-B0 classifier reached 71.74% accuracy and 70.23% macro F1 on 138 held-out images across six classes. The fitted temperature of 1.030532 indicated that the raw scores were already near calibrated, and an end-to-end run with a real model returned Atopic Dermatitis at 82.52% confidence. Isolating the model behind a service boundary keeps the system maintainable as the model changes.

The work is a prototype. Its accuracy suits screening and awareness but not clinical reliance, its severity grade is a visual cue rather than a clinical score, and its fusion model awaits data. Future work, in order of priority, is as follows.

- Data: collect a larger, ethically approved set of AD images across skin tones, ages, body sites and cameras with expert severity labels, and report results per group.
- Retrain with AD as an explicit class and evaluate on an external, dermatologist-labelled test set.
- Clinically grounded severity: train and test an ordinal severity model against EASI or SCORAD scores.
- Multimodal fusion: train the image-plus-clinical model on paired patient-level data and compare it fairly with the image-only baseline.
- Reliability: add image-quality checks, an unknown or not-skin class, uncertainty estimation beyond a single threshold, and Grad-CAM explanations.
- Deployment and governance: containerise the service, add automated tests and continuous integration, evaluate on-device inference, and complete a clinical validation study with the required privacy and regulatory review before any public use.

IX. Acknowledgment

The authors thank their project guide, Prof. Sunil A, and the Department of Computer Science and Engineering, City Engineering College, for guidance and support throughout this work.

X. References

[1] R. J. Hay et al., “The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions,” J. Invest. Dermatol., vol. 134, no. 6, pp. 1527–1534, 2014.

[2] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[3] R. Daneshjou et al., “Disparities in dermatology AI performance on a diverse, curated clinical image set,” Sci. Adv., vol. 8, no. 31, eabq6147, 2022.

[4] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR 70, 2017, pp. 1321–1330.

[5] J. M. Hanifin et al., “The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis,” Exp. Dermatol., vol. 10, no. 1, pp. 11–18, 2001.

[6] European Task Force on Atopic Dermatitis, “Severity scoring of atopic dermatitis: the SCORAD index,” Dermatology, vol. 186, no. 1, pp. 23–31, 1993.

[7] C. H. Bang et al., “Automated severity scoring of atopic dermatitis patients by a deep neural network,” Sci. Rep., vol. 11, art. 6049, 2021.

[8] A. Medela et al., “Automatic SCOring of atopic dermatitis using deep learning: a pilot study,” JID Innovations, vol. 2, no. 3, 100107, 2022.

[9] J. Yap, W. Yolland and P. Tschandl, "Multimodal skin lesion classification using deep learning," Exp. Dermatol., vol. 27, no. 11, pp. 1261–1267, 2018.
[10] T. Baltrušaitis, C. Ahuja and L.-P. Morency, "Multimodal machine learning: a survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.
[11] P. Tschandl, C. Rosendahl and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, 180161, 2018.
[12] M. Groh et al., "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in Proc. IEEE/CVF CVPR Workshops, 2021, pp. 1820–1828.
[13] A. Widiawaty, W. Indriatmi, W. Jatmiko et al., "Multimodal machine learning approach for diagnosing atopic dermatitis," F1000Research, vol. 14, 2026.
[14] M. Lalonde and H. Ghodrati, "Vision-language models and automated grading of atopic dermatitis," arXiv:2505.17835, 2025.
[15] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), PMLR 97, 2019, pp. 6105–6114.
[16] J. Deng et al., "ImageNet: a large-scale hierarchical image database," in Proc. IEEE CVPR, 2009, pp. 248–255.
[17] A. Paszke et al., "PyTorch: an imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, 2019.
[18] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980, 2014.
[19] R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, pp. 336–359, 2020.
[20] A. Das, V. Agarwal and N. P. Shetty, "Comparative analysis of multimodal architectures for skin lesion detection," Front. Artif. Intell., 2025.

