Agentic AI-Based Intelligent Dermatological Assessment and Skin Care Decision System

Yuvraj Aute[1], Vaibhav Bobade[2], Ramnath Gadhe[3], Raviraj Avhad[4], Ashwini Shidore[5]
[1,2,3,4,5]*Artificial Intelligence and Data Science Engineering, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahilyanagar, Affiliated to Savitribai Phule Pune University, Maharashtra, India*

***Abstract*—Early and accurate assessment of skin lesions supports timely clinical care, yet access to dermatological expertise is limited and visual examination is subject to inter-observer variability. Most deep learning systems for skin lesion analysis stop at image classification and do not combine prediction reliability, risk-oriented triage and patient-facing guidance in one transparent workflow. This paper proposes an Agentic AI-based dermatological assessment and skin-care decision-support system built around a seven-agent architecture: image preprocessing, deep learning classification, confidence evaluation, risk assessment, skin-care recommendation, decision coordination and report generation. Skin lesion images are classified into the seven diagnostic categories of the HAM10000 dataset (10,015 dermatoscopic images) with an EfficientNet-B0 network initialised by transfer learning and followed by a softmax layer. The maximum softmax probability is treated as a model-confidence indicator, not as a calibrated measure of diagnostic certainty, and is combined with transparent predefined rules to recommend professional evaluation for uncertain or potentially concerning results. A coordination agent merges the agent outputs into a structured report containing the predicted category, confidence, risk level and general skin-care guidance. Only the classification agent is a learned component; the remaining agents are deterministic, rule-based software modules, and no large language model is assumed. The paper presents the design, algorithms, mathematical formulation and an evaluation framework with placeholders for measured results; no experimental performance is claimed. The system is intended for preliminary decision support and teledermatology triage under professional dermatological supervision.**

***Index Terms*—Agentic AI, Clinical decision support, Deep learning, EfficientNet-B0, HAM10000, multi-agent system, Skin lesion classification, Transfer learning**

I. INTRODUCTION

Skin disorders are among the most common reasons for seeking medical advice, and the early assessment of pigmented and non-pigmented skin lesions is central to timely treatment, particularly for malignant melanoma and non-melanoma skin cancers. In routine practice, a dermatologist inspects the lesion with the naked eye and, increasingly, with a dermatoscope, a hand-held magnifying device with polarised or immersion lighting that reveals sub-surface structures. Dermoscopy improves diagnostic accuracy compared with naked-eye examination, but the improvement depends on the training of the examiner [1]. Consequently, the quality of a first assessment varies with the availability of specialists, with the examiner's experience and with the quality of the image or the consultation.

These constraints have motivated image-based teledermatology and computer-aided diagnosis. In teledermatology, a photograph is sent to a remote clinician; this widens access but increases the number of cases that must be triaged. Artificial intelligence (AI), and in particular deep learning, has become the dominant approach to automating the interpretation of medical images. A convolutional neural network (CNN) is a deep network that learns hierarchical image features through stacked convolution and pooling operations. Esteva et al. showed that a CNN trained on a large clinical image collection could classify skin lesions at a level comparable to board-certified dermatologists on selected tasks [2], and Haenssle et al. reported a similar finding for dermoscopic melanoma recognition [3]. The public availability of the HAM10000 dataset, which contains 10,015 dermatoscopic images in seven diagnostic

categories [4], has since made reproducible multi-class benchmarking possible.

EfficientNet-B0 is the baseline member of the EfficientNet family. It uses mobile inverted bottleneck (MBConv) blocks with squeeze-and-excitation and was found by a neural architecture search; larger family members are obtained by jointly scaling depth, width and input resolution with a compound coefficient [5]. With roughly 5.3 million parameters, EfficientNet-B0 is compact enough for web and mobile deployment while retaining the benefits of ImageNet pre-training, which makes it a suitable backbone for a lightweight decision-support service.

Most published skin-lesion systems, however, end at the classifier: they output a class label or a probability vector. A clinician or a patient needs more than that. The reliability of the prediction must be judged, the result must be translated into an action (for example, whether professional evaluation is advisable), and the output must be communicated in a structured way. Agentic AI addresses this kind of workflow. An agent is a software entity that is situated in an environment, acts autonomously within a defined scope, and interacts with other agents to achieve a goal [6]; multi-agent organisations of this kind are increasingly examined for clinical decision support [7].

This paper proposes an Agentic AI-based dermatological assessment and skin-care decision system. Its contributions are as follows.

- A seven-agent architecture that separates image preprocessing, classification, confidence evaluation, risk assessment, skin-care recommendation, decision coordination and report generation, each with a defined interface.
- An EfficientNet-B0 classifier, adapted by transfer learning, for the seven HAM10000 lesion categories.
- A confidence-aware and rule-based risk-assessment stage that is transparent and auditable, and that escalates uncertain or potentially concerning results to professional evaluation.
- A precise description of the implementation of the multi-agent workflow, in which only the classifier is a learned component and the other agents are deterministic software modules.
- Seven algorithms with mathematical formulation and an evaluation framework that is appropriate for the strong class imbalance of HAM10000.

The system is intended for preliminary decision support only. It is not an autonomous diagnostic device and does not replace professional dermatological evaluation. This paper specifies the design and the evaluation protocol; the sections that report measured performance contain clearly marked placeholders, and no experimental result is claimed. The rest of the paper is organised as follows. Section II reviews existing systems, Section III reviews the literature, and Section IV compares the approaches. Sections V–VIII describe the proposed system, its methodology, algorithms and architecture. Sections IX and X present the evaluation framework and the results template, and Sections XI–XIII discuss applications, limitations and conclusions.

II. EXISTING SYSTEM

The approaches that are currently used for skin lesion assessment are summarised below, together with their main advantages and limitations. A schematic of the conventional workflow is given in Fig. 1.

A. Traditional Dermatological Examination

The reference workflow consists of clinical history, visual inspection, dermoscopy and, when the lesion is suspicious, excision biopsy with histopathological confirmation. Its main advantage is that the clinician integrates the lesion's appearance with patient history and context, and histopathology gives a definitive diagnosis. Its limitations are the dependence on scarce specialist time, inter-observer variability, and the fact that accuracy with dermoscopy depends on examiner training [1]. It also produces no structured, reusable record unless one is created manually.

B. Machine Learning-Based Classification

Classical computer-aided systems extract hand-crafted colour, texture and shape descriptors from a (usually segmented) lesion and pass them to a classifier such as a support vector machine; Celebi et al., for example, combined feature selection with a support vector machine for dermoscopy images [8]. These methods are lightweight and relatively easy to interpret. Their performance, however, depends on the quality of segmentation and on feature engineering, and they scale poorly to the visual variability of multi-class problems.

C. CNN-Based Skin Lesion Classification

CNN-based systems learn features directly from pixels and have reached dermatologist-level performance on specific tasks and test sets [2], [3]. Their advantage is accuracy and automation of feature learning. Their limitations are the need for large labelled datasets, sensitivity to data distribution, limited generalisation to image sources that were not seen during training [9], and the lack of any mechanism beyond the predicted label.

D. Transfer Learning Approaches

Transfer learning reuses knowledge from a source task to improve learning on a target task with less data [10]. In dermatology, networks pre-trained on ImageNet [11] are fine-tuned on dermoscopic images; this shortens training and generally improves accuracy on small medical datasets. The limitation is domain shift between natural and dermoscopic images, and the risk of over-fitting when fine-tuning on a small or imbalanced dataset.

E. Explainable AI Systems

Explainable AI (XAI) techniques such as Grad-CAM [12] highlight the image regions that influenced a prediction. Systems such as SkinSage XAI combine a CNN with such visual explanations to support clinician trust [13]. XAI improves transparency, but saliency maps are post-hoc and do not by themselves tell the user how reliable a prediction is or what action to take.

F. Ensemble Deep Learning

Ensembles combine the outputs or features of several networks. DCENSnet is a representative example that fuses three deep convolutional networks for skin cancer classification [14]. Ensembles often improve robustness and accuracy, but multiply memory use and inference time, which is a drawback for low-resource or mobile deployment.

G. Segmentation and Classification Approaches

Two-stage or dual-branch models first delineate the lesion and then classify it, using appearance and morphological features of the segmented region [15]. They offer lesion-level localisation and can improve classification, but they require pixel-level masks for training, add architectural complexity, and the classifier is affected by segmentation errors.

H. Clinical Decision-Support Systems

Clinical decision-support systems (CDSS) deliver patient-specific recommendations to clinicians. Recent work studies multi-agent CDSS built on reinforcement learning, ontologies and large language models [7]. Human–computer collaboration studies suggest that how AI output is presented to clinicians affects its benefit [16]. The limitation is that most multi-agent CDSS work is general-purpose, and little of it addresses image-based dermatological triage with transparent, auditable rules.

III. LITERATURE REVIEW

This section reviews eight studies that frame the proposed system. Quantitative results below are those reported by the cited authors on their own data, and they are not results of the proposed system.

A. Esteva et al. (2017)

Objective and method. Esteva et al. investigated whether a single deep CNN could classify skin lesions as accurately as dermatologists [2]. They fine-tuned an Inception-v3 network, pre-trained on ImageNet, using a dataset of 129,450 clinical images covering 2,032 diseases. Contribution. The network was tested against 21 board-certified dermatologists on biopsy-proven cases for keratinocyte carcinoma versus benign seborrheic keratosis and for malignant melanoma versus benign nevus, and it matched dermatologist performance on those binary tasks. Limitations and relevance. The tasks were binary and the evaluation used curated images; the study did not provide a triage or reporting workflow. It established the viability of transfer learning with CNNs for skin imagery, which underpins our choice of an ImageNet-pre-trained backbone.

B. Tschandl et al. (2018)

Objective and dataset. Tschandl, Rosendahl and Kittler released HAM10000 (“Human Against Machine with 10000 training images”), a collection of 10,015 dermatoscopic images of pigmented lesions from two sources, to provide diverse, publicly available training data [4]. The dataset covers seven diagnostic categories (MEL, NV, BCC, AKIEC, BKL, DF and VASC), and the diagnoses were confirmed by histopathology in more than half of the cases, with the others established by follow-up, expert consensus or

confocal microscopy. Limitations and relevance. The dataset is strongly imbalanced, covers only seven pigmented-lesion categories rather than all skin diseases, and contains several images of some lesions, so splits must be made by lesion. It is the training and evaluation dataset for the proposed classifier, and it was the basis of the ISIC 2018 classification task [17].

C. Tschandl et al. (2020)
Objective and method. This study examined how AI decision support should be presented to clinicians [16]. A reader study with 302 raters from 41 countries, including board-certified dermatologists, residents and general practitioners, compared different representations of CNN output. Contribution. Good-quality AI support improved diagnostic accuracy over that of clinicians alone, the least experienced clinicians gained the most, and the form in which the AI output was presented influenced the benefit. Limitations and relevance. The study used a reader-study design rather than routine care. Its findings motivate our design choices to present probabilities and confidence together with a risk statement, and to keep a clinician in the loop.

D. Daneshjou et al. (2021)
Objective and method. Daneshjou et al. performed a scoping review of 70 studies that developed AI algorithms for skin disease, auditing how the datasets were described [18]. Contribution. They found sparse dataset characterisation and transparency, non-standard and unverified disease labels, and an inability to assess the diversity of patients, including skin tone, with only a small minority of the reviewed studies reporting any skin-tone information. Of about 1.07 million images, only about a quarter were publicly available or available on request. Limitations and relevance. It is a review and does not propose a model. It directly informs our discussion of dataset bias, the representation of skin tones in HAM10000, and the insistence on transparent reporting.

E. Munjal et al. (2024)
Objective and method. SkinSage XAI is an explainable framework for seven-class skin lesion diagnosis [13]. It classifies images with an Inception-v3 network and explains predictions with Grad-CAM and LIME (local interpretable model-agnostic explanations). Dataset and results. The authors used a customised dataset of about 50,000 images derived from HAM10000 and reported 96% accuracy, 96.42% precision, 96.28% recall, 96.14% F1-score and an AUC of 99.83%. Limitations and relevance. The explanations are post-hoc visualisations, evaluation is reported on a single HAM10000-derived dataset, and no reliability-aware triage or report generation is described. The study confirms the value of transparency; our system supplements transparency with a confidence stage and explicit action rules.

F. Chanda et al. (2024)
Objective and method. DCENSnet is a deep convolutional ensemble network for skin cancer classification that combines three deep CNNs to improve robustness over single models [14]. Contribution. It illustrates that ensembling can raise accuracy for skin cancer classification. Limitations and relevance. Running three networks raises computation and memory requirements, the work addresses classification only, and its performance figures are specific to its own data and protocol. We use a single compact backbone, EfficientNet-B0, and obtain additional functionality from the agent layer rather than from a larger ensemble.

G. Abu Owida et al. (2025)
Objective and method. Abu Owida et al. developed a dual-branch framework that performs lesion segmentation and classification in dermoscopic images [15]. The segmentation branch uses a modified EfficientNet-B7 encoder with atrous spatial pyramid pooling, transformer blocks, attention gates and squeeze-and-excitation blocks; the classification branch fuses DenseNet-121 features with morphological features extracted from the predicted masks. Dataset and results. Experiments covered five benchmark datasets including HAM10000, PH2 and ISIC 2016–2018; on HAM10000 the authors report a Dice score of 0.9568, an IoU of 0.9242 and an accuracy of 0.9708. Limitations and relevance. The architecture is computationally heavy, classification depends on segmentation quality, and there is no risk assessment or user-facing guidance. Our design takes the opposite trade-off: a lightweight classifier and a rule-based decision layer.

H. Silveira et al. (2026)
Objective and method. Silveira, da Rosa Righi and da Costa conducted a PRISMA-guided systematic review of 42 studies published between 2016 and 2025 on

multi-agent reinforcement learning, ontology-based knowledge representation and generative AI in clinical decision support [7]. Contribution. They propose a taxonomy and identify gaps in hierarchical architectures, bidirectional knowledge flow and explainability. Limitations and relevance. As a review, it provides no new system, and it is not specific to dermatology. It supports modular, role-specialised agent architectures for clinical decision support. Our system uses a much simpler coordination mechanism, and does not claim reinforcement learning, ontologies or LLM reasoning.

I. Summary of the Gap

The reviewed work shows strong classifiers [2], [13], [14], [15], rich datasets [4], evidence on human–AI interaction [16], caution about dataset bias [18], and a growing interest in multi-agent decision support [7]. What is lacking is a compact, transparent pipeline that links a lightweight classifier on HAM10000 to explicit confidence handling, rule-based risk assessment and structured reporting under human supervision. The proposed system targets this gap.

IV. COMPARATIVE ANALYSIS

Table I compares the approaches discussed in Sections II and III. The entries are qualitative characterisations drawn from the reviewed literature and are not the outcome of a head-to-head experiment.

Table I. Comparative Analysis of Approaches

Approach	Advantages	Limitations
Manual dermatological assessment	Contextual judgement; histopathological confirmation; accepted clinical standard	Limited specialist access; inter-observer variability; unstructured record
Traditional machine learning	Light computation; interpretable hand-crafted features	Depends on segmentation and feature engineering; limited multi-class scalability
CNN-based classification	Automatic feature learning; expert-level accuracy on some tasks	Needs large data; poor out-of-distribution robustness; label-only output
Transfer learning models	Faster convergence; effective with limited data	Domain shift; over-fitting on small or imbalanced sets
Ensemble deep learning	Higher robustness and accuracy	Several networks: higher memory and latency
Explainable AI models	Visual evidence supports trust	Post-hoc maps; no reliability or action guidance
Segmentation and classification models	Lesion localisation; morphology-aware features	Needs masks; error propagation; heavy architecture
Multi-agent clinical decision-support models	Modular roles; coordination; extensible	Mostly general-purpose; limited clinical validation and dermatology-specific work
Proposed Agentic AI-based dermatological assessment system	Compact EfficientNet-B0; confidence-aware; transparent rules; structured report; modular agents	Seven HAM10000 classes only; uncalibrated confidence; thresholds need tuning; not clinically validated

Key observations. First, accuracy-oriented research (CNNs, transfer learning, ensembles, segmentation-based models) has progressed quickly, but accuracy gains typically come with greater computation or data requirements. Second, transparency research (XAI) tells the user where the model looked, not how much to trust the result or what to do next. Third, human–computer collaboration is influenced by how AI output is presented [16], so a classifier alone is an incomplete product. Fourth, dataset transparency and bias remain unresolved issues [18], which means that any system built on HAM10000 should state its scope explicitly.

How the proposed system addresses these limitations. The proposed system keeps the classifier compact (a single EfficientNet-B0), which limits latency and memory. It adds an explicit confidence-evaluation agent and a rule-based risk-assessment agent, so that low-confidence or potentially concerning outputs are routed to professional evaluation rather than being presented as definitive. It produces a structured report that records the prediction, confidence, risk level and general guidance in a consistent form, and it keeps a clinician in the loop. These advantages are design

properties; whether they translate into measurable benefit has to be established by the evaluation described in Section IX, and no superiority over any existing system is claimed here.

V. PROPOSED SYSTEM

A. Overall System Architecture

The proposed system is organised in five layers, as shown in Fig. 2: a user-interface layer for image upload and report display, an agent-orchestration layer, a specialised-agent layer, a model layer containing the EfficientNet-B0 classifier, and a knowledge layer that stores the risk rules and guidance templates. Data flow between agents is described in Fig. 4 and the end-to-end workflow in Fig. 3. Throughout the paper, agents are denoted A1–A7.

B. System Components

Table II lists the seven agents. Each agent has a single responsibility, a typed input and a typed output, so that an agent can be modified or replaced without changing the others.

Table II. Agents and Their Responsibilities

Agent	Responsibility and output	Implementation
A1 Image Preprocessing Agent	Checks, resizes and normalises the uploaded image; outputs an image tensor	Deterministic module
A2 Deep Learning Classification Agent	Runs EfficientNet-B0; outputs logits and softmax probabilities	Learned model
A3 Confidence Evaluation Agent	Computes confidence indicators and sets an uncertainty flag	Deterministic module
A4 Risk Assessment Agent	Applies predefined rules; outputs a risk level and a recommendation	Rule-based
A5 Skin-Care Recommendation Agent	Selects general, non-prescriptive guidance	Template / rule-based
A6 Decision Coordination Agent	Sequences the agents, handles exceptions and merges outputs	Workflow controller
A7 Report Generation Agent	Assembles the structured assessment report	Template-based

Implementation note. The terms “agent” and “multi-agent” are used in the software-engineering sense: each agent is an autonomous module with its own goal, state and interface, and agents cooperate through a shared case record [6]. Only A2 is a learned model. A1, A3, A4, A5 and A7 are deterministic modules, and A6 is a deterministic workflow controller. They are not independent reasoning agents, and no large language model (LLM) is used or assumed. An LLM could be added later, for example to phrase report text, but this would be an extension and not part of the present design.

C. Input Image Processing

The user uploads a dermatoscopic or close-up skin image. A1 verifies the format, rejects files that cannot be decoded or that are far too small, resizes the image to 224 × 224 pixels, scales the intensities and normalises each colour channel with the statistics that match the pre-trained weights. The output is a tensor that is passed to A2 through the shared record.

D. EfficientNet-B0 Classification

A2 applies the fine-tuned EfficientNet-B0 to the tensor and returns seven logits, which are converted into class probabilities by a softmax layer. The classes are the HAM10000 categories: MEL (melanoma), NV (melanocytic nevi), BCC (basal cell carcinoma), AKIEC (actinic keratoses and intraepithelial carcinoma), BKL (benign keratosis-like lesions), DF (dermatofibroma) and VASC (vascular lesions) [4]. The classifier is therefore applicable only to these seven categories and does not cover other skin diseases.

E. Multi-Agent Communication

Agents do not call one another directly. A6 creates a case record, a typed data structure holding the image identifier, tensor, probabilities, predicted class, confidence indicators, uncertainty flag, risk level, guidance and report fields. A6 invokes each agent in a

fixed order, passes it the case record, validates the fields that the agent returns and writes them back. Because every message is a named field, the full decision trail can be logged and audited. The control flow is a pipeline with one conditional branch (uncertain versus confident predictions), which keeps behaviour predictable, an important property in a medical setting.

F. Confidence Evaluation

A3 uses the maximum softmax probability, the margin between the two largest probabilities and, optionally, the normalised entropy. The maximum softmax probability is a model-confidence indicator, not a calibrated measure of diagnostic certainty: deep networks are often over-confident [19]. A3 therefore only marks a prediction as uncertain when the indicators fall below configured thresholds, and does not present the confidence value as a probability of disease.

G. Rule-Based Risk Assessment

A4 maps the predicted class, the confidence indicators and the probability mass of the potentially concerning classes to one of four risk levels, using the transparent rules of Table IV (Section VII). The rules are conservative: uncertainty, or an appreciable probability of MEL, BCC or AKIEC, leads to a recommendation of professional evaluation, even when the top-1 class is benign. The rules are decision-support rules and not clinical guidelines.

H. Skin-Care Recommendation

A5 selects general guidance from a fixed template library: broad-spectrum sun protection and avoidance of tanning, gentle cleansing and moisturising, avoiding picking or irritating a lesion, photographing and monitoring a lesion for change (the ABCDE features: asymmetry, border irregularity, colour variation, diameter and evolution), and seeking professional advice for new, changing, bleeding or painful lesions. A5 does not recommend or dose medication and does not give treatment instructions.

I. Final Report Generation

A7 produces a structured report with the following fields: case identifier and timestamp; image-quality status; predicted category with the full probability vector; confidence indicators and uncertainty flag; risk level and rationale (the rule that fired); general skin-care guidance; and a standard statement that the report is not a diagnosis and that professional dermatological evaluation is necessary for medical decisions.

J. Human Clinical Supervision

The report is addressed to a clinician or to a user who is advised to consult a clinician. The system never closes the loop on its own: a recommended action is always “seek professional evaluation” or “monitor and seek evaluation if the lesion changes”. This design reflects evidence that AI is most useful in collaboration with clinicians [16].

K. End-to-End Workflow

The workflow of Fig. 3 is as follows. (1) The user uploads an image. (2) A1 checks and preprocesses it. (3) A2 computes the class probabilities. (4) A3 evaluates confidence and sets the uncertainty flag. (5) A4 assigns a risk level; if the prediction is uncertain, the result is escalated to “professional evaluation recommended”. (6) A5 selects general guidance. (7) A6 merges the outputs and checks consistency. (8) A7 renders the structured report. (9) The report is reviewed by the user or clinician.

VI. SYSTEM METHODOLOGY

The methodology consists of fourteen stages. The training and evaluation settings described below are the proposed protocol; the final values are to be fixed and reported together with the measured results.

A. Dataset Collection

The HAM10000 dataset provides 10,015 dermatoscopic images of pigmented lesions together with metadata (diagnosis, age, sex, anatomical site and a lesion identifier) [4]. It was assembled from two sources over a long period, and diagnoses were confirmed by histopathology, follow-up, expert consensus or confocal microscopy. It covers seven categories and does not include every skin disease.

B. Dataset Exploration and Class Distribution

The class distribution is shown in Table III. The dataset is highly imbalanced: NV accounts for about two-thirds of the images, while DF and VASC together account for about 2.6%. A classifier trained naively tends to favour NV, and overall accuracy then

conceals poor performance on rare but clinically important classes. Exploration should also include the distributions of age, sex and body site, and the number of images per lesion.

Table III. HAM10000 Class Distribution

Code	Category	Images	%
NV	Melanocytic nevi	6,705	66.95
MEL	Melanoma	1,113	11.11
BKL	Benign keratosis-like	1,099	10.97
BCC	Basal cell carcinoma	514	5.13
AKIEC	Actinic keratoses / intraepithelial ca.	327	3.27
VASC	Vascular lesions	142	1.42
DF	Dermatofibroma	115	1.15
Total		10,015	100.00

C. Image Preprocessing

Images are resized to the EfficientNet-B0 input resolution of 224 × 224 pixels, converted to three-channel RGB, scaled to [0, 1] and normalised channel-wise using the mean and standard deviation of the pre-training data. Preprocessing in training and in deployment must be identical; this is enforced by A1 using the same function in both settings (Algorithm 1).

D. Data Augmentation

Augmentation is applied to the training set only. Random flips, rotations, small zooms and mild brightness and contrast changes increase the effective size of the data and reduce over-fitting, in particular for minority classes. Augmentation must not change the diagnostic content of the image, so strong colour distortions are avoided (Algorithm 2).

E. Training and Validation Dataset Preparation

A stratified split into training, validation and test subsets is proposed (for example 70/15/15). Because HAM10000 contains multiple images of some lesions, the split must be made by lesion identifier so that images of the same lesion never appear in more than one subset; otherwise, the reported performance is optimistic. The test subset is used once, after all design decisions have been made.

F. EfficientNet-B0 Model Initialisation

The network is initialised with ImageNet-pre-trained weights [11], [5], and the original 1,000-way classifier is replaced by a dropout layer followed by a dense layer with seven outputs. Transfer learning of this kind reuses generic low-level and mid-level features and usually helps when the target dataset is small [10].

G. Feature Extraction

The convolutional backbone (stem, seven stages of MBConv blocks and a 1 × 1 convolution head) maps the image to a feature map, and global average pooling reduces it to a 1,280-dimensional feature vector. In the first training phase the backbone is frozen and acts as a fixed feature extractor.

H. Model Training and Classification

Training proceeds in two phases: the new classifier is trained with the backbone frozen, and then the backbone is unfrozen and fine-tuned at a lower learning rate. A weighted cross-entropy loss, or alternatively oversampling of minority classes, addresses class imbalance. The Adam optimiser [20], a batch size of 32 and early stopping on validation macro-averaged F1-score are proposed starting points. Model selection uses the validation set only.

I. Softmax Probability Calculation

The seven logits are converted to a probability vector by the softmax function. The predicted class is the argument of the maximum probability (Algorithm 4).

J. Confidence Evaluation

A3 derives the confidence indicators from the probability vector and sets the uncertainty flag. Thresholds are selected on the validation set, for example to reach a target sensitivity for the concerning classes, and are then frozen before testing (Algorithm 5).

K. Risk Assessment

A4 applies the rules of Table IV, which depend on the predicted class, the uncertainty flag and the probability mass of the potentially concerning classes (Algorithm 6).

L. Multi-Agent Coordination

A6 executes the pipeline, checks the validity of each agent's output, handles failures (for example an unreadable image) by terminating the pipeline with an explanatory report, and logs the case record (Algorithm 7).

M. Skin-Care Recommendation

A5 selects the guidance text associated with the risk level from the template library and appends the standard precautions. No treatment or dosage is generated.

N. Final Assessment Report

A7 renders the report fields in a fixed layout and includes the disclaimer. The report can be displayed in the interface or exported as a document.

VII. ALGORITHMS

This section specifies the seven algorithms. Let K = 7 be the number of classes, x the uploaded image, yi the i-th class and z = (z1, …, zK) the vector of logits.

Algorithm 1: Image Preprocessing

Input: Uploaded image x, target size S = 224, channel means μc and standard deviations σc

Output: Normalised tensor xn, or an error status

1. if x cannot be decoded or min (H, W) < s_{min} then return ERROR
2. Convert x to RGB
3. Resize x to S × S (bilinear interpolation)
4. for each channel c and pixel (i, j) do
5. u_c (i, j) ← x_c (i, j) / 255
6. $x_{n,c}$(i, j) ← (u_c (i, j) − μ_c) / σ_c
7. end for
8. return x_n

Algorithm 1 enforces a uniform input format. The scaling and normalisation of each channel are

uc (i, j) = xc (i, j) / 255 (1)

xn,c(i, j) = (uc (i, j) − μc) / σc (2)

where c indexes the colour channel and (i, j) the pixel. The values of μc and σc must match the convention of the pre-trained weights (for ImageNet weights, the ImageNet channel statistics). The minimum size smin is a quality threshold, not a diagnostic criterion.

Algorithm 2: Data Augmentation (training only)

Input: Training image xn, label y, parameters θmax, pflip, zoom range zr, jitter strength δ

Output: Augmented image xa with unchanged label y

1. xa ← xn
2. with probability pflip: flip xa horizontally
3. with probability pflip: flip xa vertically
4. Draw θ ~ U (−θmax, θmax); rotate xa by θ
5. Draw s ~ U (1 − zr, 1 + zr); zoom xa by s and crop to S × S
6. Draw b, k ~ U (1 − δ, 1 + δ); xa ← clip (k (xa − m) + m · b, 0, 1)
7. return (xa, y)

Rotation by angle θ maps a pixel coordinate according to

(x′, y′) T = R(θ) (x, y) T (3)

R(θ) = [cos θ, −sin θ; sin θ, cos θ] (4)

In step 6, m is the mean intensity of the image, b the brightness factor and k the contrast factor. Class imbalance is handled in the loss by class weights

ωk = N / (K nk) (5)

where N is the number of training images and nk the number in class k; rare classes thus receive larger weights. Augmentation is never applied to the validation or test subsets.

Algorithm 3: EfficientNet-B0 Classification

Input: Normalised tensor xn (224 × 224 × 3), pre-trained weights Θ0

Output: Logit vector z ∈ RK

1. if training: load Θ0; replace classifier by Dropout + Dense (1280 → K)
2. F ← Stem(xn) (3 × 3 convolution, stride 2, batch norm, Swish)
3. for stage s = 1 to 7 do
4. F ← MBConv blocks of stage s(F) (depthwise convolution, squeeze-and-excitation)
5. end for
6. F ← Conv1×1(F) with 1,280 channels, batch norm, Swish
7. f ← GlobalAveragePool(F) (f ∈ R1280)
8. z ← WTf + η
9. return z

The network uses the Swish activation and a linear classification layer:

swish(x) = x · σ(x), σ(x) = 1 / (1 + exp(−x)) (6)

z = WTf + η (7)

where W ∈ R1280×K and η ∈ RK are the weights and bias of the new classifier. EfficientNet-B0 is the baseline of a family that is scaled with a compound coefficient φ, with depth d = αφ, width w = βφ and resolution r = γφ, subject to α · β2 · γ2 ≈ 2 [5]; B0 corresponds to φ = 0, and no scaling is applied here. During fine-tuning the weighted cross-entropy loss is minimised:

L = − (1/N) Σn Σk ωk yn, k log pn, k (8)

where yn, k is the one-hot label and pn, k the softmax probability of class k for image n.

Algorithm 4: Softmax Probability Calculation

Input: Logit vector z, optional temperature T ≥ 1 (default T = 1)

Output: Probability vector P, predicted class ŷ, confidence C

1. m ← maxj zj (for numerical stability)
2. for i = 1 to K do ei ← exp ((zi − m) / T)
3. Z ← Σj ej
4. for i = 1 to K do Pi ← ei / Z
5. ŷ ← argmaxi Pi
6. C ← maxi Pi
7. return (P, ŷ, C)

The softmax function converts logits to a probability distribution over the K classes:

$$P(y_i \mid x) = \exp(z_i) / \Sigma_j \exp(z_j) \quad (9)$$

The predicted class and the confidence score are

$$\hat{y} = \text{argmax}_i P(y_i \mid x), \; C = \max_i P(y_i \mid x) \quad (10)$$

Subtracting m does not change the result because the softmax is invariant to a constant shift of the logits. Temperature scaling (T > 1) can be fitted on the validation set to reduce over-confidence [19]. C is a model-confidence indicator. It is not necessarily a calibrated measure of diagnostic certainty, and a high C does not guarantee a correct diagnosis.

Algorithm 5: Confidence Evaluation

Input: Probability vector P, thresholds Tconf, Tm, (optional) TH

Output: Indicators (C, m, Hn) and uncertainty flag u ∈ {0, 1}

1. Sort P in descending order: P (1) ≥ P (2) ≥ …
2. C ← P (1); m ← P (1) − P (2)
3. Hn ← − (Σi Pi log Pi) / log K
4. if C < Tconf or m < Tm or Hn > TH then u ← 1
5. else u ← 0
6. return (C, m, Hn, u)

In addition to the maximum probability C, the margin m and the normalised entropy

$$m = P(1) - P(2), \; Hn = -(1/\log K) \Sigma_i P_i \log P_i \quad (11)$$

describe how decisively the network prefers one class (Hn ∈ [0, 1]). The thresholds are tuned on the validation set and are design parameters; the value Tconf = 0.60 is used only as an illustrative starting point. Their effect should be assessed with calibration measures such as the expected calibration error given in Section IX.

Algorithm 6: Rule-Based Risk Assessment

Input: Predicted class ŷ, probability vector P, uncertainty flag u, quality status q, threshold Tconc

Output: Risk level r and recommendation

1. Let G = {MEL, BCC, AKIEC}; Sconc ← Σk∈G Pk
2. if q = FAIL then return (RECAPTURE, “Upload a clearer image”)
3. if u = 1 then r ← UNCERTAIN
4. else if ŷ ∈ G then r ← HIGH CONCERN
5. else if Sconc ≥ Tconc then r ← ELEVATED
6. else r ← LOWER CONCERN
7. return (r, recommendation(r))

The probability mass of the concerning classes G = {MEL, BCC, AKIEC} is

$$Sconc = \Sigma_{k \in G} P(y_k \mid x) \quad (12)$$

The rules and their recommendations are listed in Table IV. The set G contains the three categories that are malignant or pre-malignant. The safety-net rule on Sconc routes a case to professional evaluation even when a benign class has the highest probability. The thresholds are design parameters to be tuned on validation data (for example Tconc = 0.20 as a starting value) and are not clinically validated. Even the “LOWER CONCERN” level never states that a lesion is harmless.

Table IV. Rule-Based Risk Assessment (evaluated in order)

Rule	Condition	Risk level	Recommendation
R1	Image quality fails	RECAPTURE	Upload a clearer, well-lit image
R2	u = 1	UNCERTAIN	Professional evaluation recommended
R3	ŷ ∈ {MEL, BCC, AKIEC}	HIGH CONCERN	Prompt professional evaluation
R4	Sconc ≥ Tconc	ELEVATED	Professional evaluation advised
R5	Otherwise	LOWER CONCERN	Monitor (ABCDE); evaluation if it changes

Algorithm 7: Multi-Agent Coordination

Input: Uploaded image x; agents A1–A5, A7; configuration τ (thresholds, rule set)

Output: Report R (or error report)

1. Create case record S0 ← {id, x, timestamp, τ}
2. S ← A1(S0); if S. status = ERROR then return ErrorReport(S)

3. S ← A2(S); validate that P is a probability vector
4. S ← A3(S)
5. S ← A4(S)
6. S ← A5(S)
7. Check consistency (risk level = rule output; guidance matches risk level)
8. R ← A7(S); write S to the audit log
9. return R

A6 composes the agents sequentially; each agent reads fields from the case record and adds its own. Writing S0 for the initial record and Ai for the function realised by the i-th agent in the order of Algorithm 7,

$$S_i = S_{i-1} \cup A_i(S_{i-1}),\ R = A_7(S_5) \quad (13)$$

where ∪ denotes the addition of the new fields. If any validation step fails, the pipeline stops and an error report is returned; no partial or inconsistent assessment is shown. The deterministic order and the logged record make every decision reproducible, which would be harder with free-form communication among agents.

VIII. SYSTEM ARCHITECTURE AND FIGURES

This section presents the five figures that describe the existing workflow and the proposed system. All figures are monochrome flowcharts.

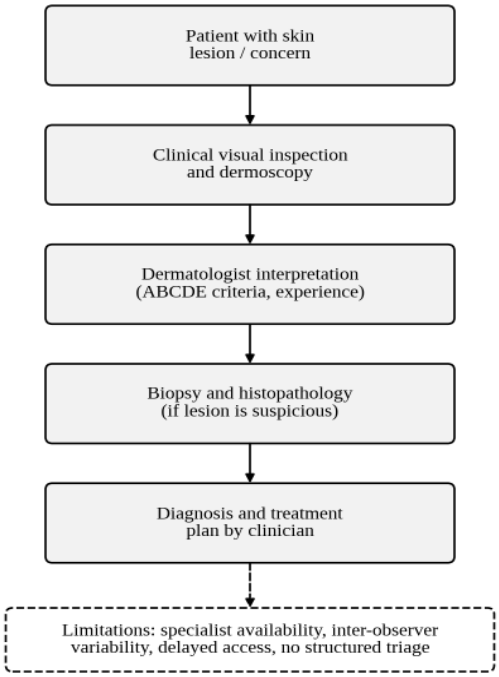


Fig. 1. Architecture of the existing dermatological assessment system.

Explanation. Fig. 1 shows the conventional pathway. A patient presents with a lesion; the clinician performs visual inspection and dermoscopy; the dermatologist interprets the findings using criteria such as ABCDE and experience; suspicious lesions undergo biopsy and histopathology; and a diagnosis and treatment plan result. Each box depends on the preceding one, so delays at any step delay the whole pathway. The dashed box summarises the limitations: dependence on specialist availability, inter-observer variability, delayed access and the absence of structured triage [1].

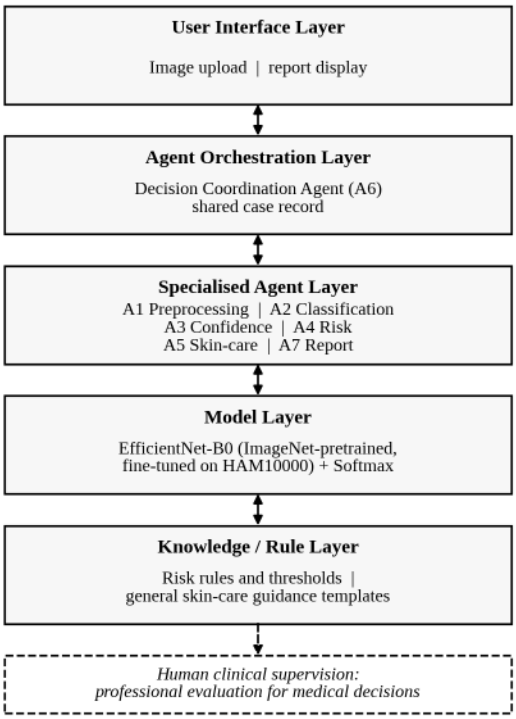


Fig. 2. Proposed system architecture.

Explanation. Fig. 2 shows five layers. The user-interface layer accepts the image and displays the report. The agent-orchestration layer holds A6, which owns the shared case record. The specialised-agent layer contains A1–A5 and A7. The model layer contains the fine-tuned EfficientNet-B0 and its softmax output and is called only by A2. The knowledge layer stores the risk rules, thresholds and guidance templates used by A3–A5. Double-headed arrows denote request/response exchanges between adjacent layers. The dashed box shows that human clinical supervision follows the system, and that the report supports but does not replace professional evaluation.

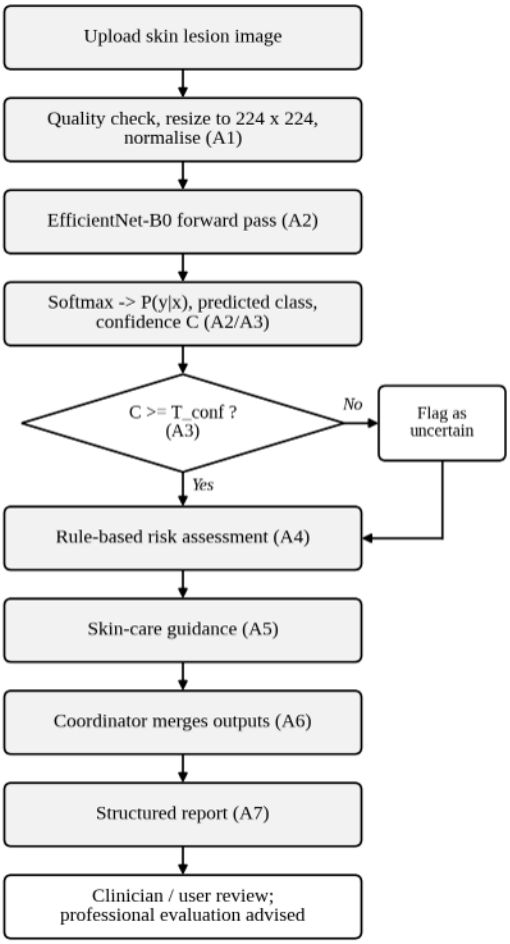


Fig. 3. Proposed system workflow.

Explanation. Fig. 3 follows one case from upload to report. The image is checked and normalised (A1), classified (A2) and converted to probabilities and a confidence score (A2/A3). A decision node tests whether the confidence reaches the threshold Tconf. If not, the prediction is flagged as uncertain, and the flag is carried into the risk-assessment step (A4), which then escalates the case to professional evaluation; otherwise, the rules of Table IV are applied normally. Guidance is selected (A5), the coordinator merges all outputs (A6), the report is generated (A7) and the user or clinician reviews it, with professional evaluation advised.

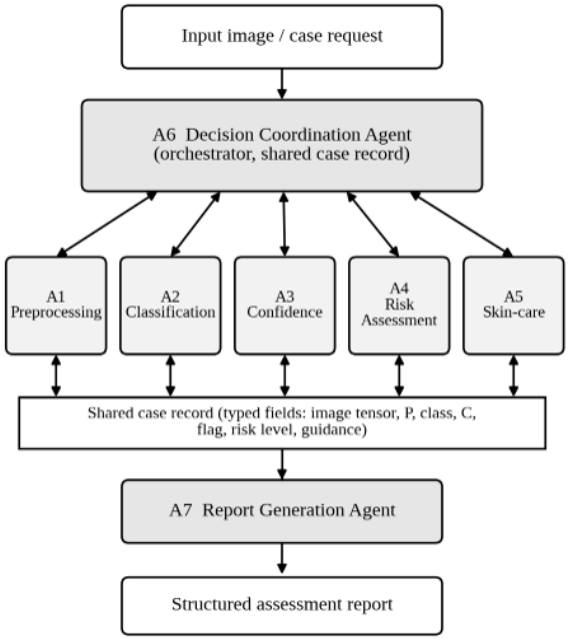


Fig. 4. Multi-agent coordination architecture.

Explanation. In Fig. 4, A6 receives the case request and invokes A1–A5 in sequence. Each agent exchanges data with the shared case record, shown as the horizontal box, by reading the fields it needs and writing its results; arrows between the agents and the record are double-headed for this reason. A6 monitors the record, validates each output and, when the record is complete, triggers A7, which renders the structured assessment report. The agents never communicate with each other directly, which makes the decision trail auditable and each agent independently testable.

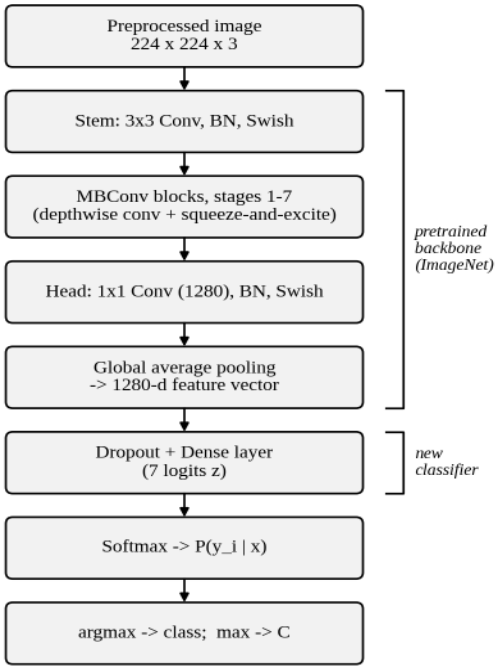


Fig. 5. EfficientNet-B0-based classification workflow.

Explanation. Fig. 5 shows the classifier of A2. A 224 × 224 × 3 tensor passes through the stem convolution, the seven stages of MBConv blocks, a 1 × 1 convolution head and global average pooling, producing a 1,280-dimensional feature vector. The stem, MBConv stages and head form the pre-trained backbone (bracket), initialised from ImageNet weights [5]. A new dropout and dense layer map the feature vector to seven logits, the only component that is trained from scratch, followed by softmax to give class probabilities. Taking the argmax and the maximum gives the predicted class and the confidence C defined in (9) and (10).

IX. MATHEMATICAL MODEL AND PERFORMANCE EVALUATION

This section defines the evaluation framework. It contains no measured values; the values are to be obtained from the implementation and reported in Section X.

A. Class Imbalance and Evaluation Protocol

In HAM10000 the majority class (NV) contains about 6,700 of 10,015 images, while the smallest class (DF) contains 115 (Table III). A classifier that always predicts NV would therefore reach an accuracy of about 67% while missing every melanoma. Overall accuracy alone is thus inadequate. Evaluation should report per-class metrics, macro-averaged metrics (which weight every class equally and expose weak minority classes) and weighted-averaged metrics (which reflect the class frequencies), together with the confusion matrix. All metrics are computed on a held-out test subset that is separated by lesion identifier and that was not used for model selection.

B. Classification Metrics

For class k, let TPk, FPk, FNk and TNk be the numbers of true positives, false positives, false negatives and true negatives in a one-versus-rest view, nk the number of test images of class k, and N the total. Accuracy is the fraction of correct predictions:

$$\mathrm{Acc} = (1/N)\ \Sigma_n\ 1[\hat{y}_n = y_n] \quad (14)$$

Precision, recall (sensitivity), specificity and F1-score of class k are, respectively,

$$\mathrm{Prk} = \mathrm{TPk} / (\mathrm{TPk} + \mathrm{FPk}) \quad (15)$$

$$\mathrm{Rek} = \mathrm{TPk} / (\mathrm{TPk} + \mathrm{FNk}) \quad (16)$$

$$\mathrm{Spk} = \mathrm{TNk} / (\mathrm{TNk} + \mathrm{FPk}) \quad (17)$$

$$\mathrm{F1k} = 2\ \mathrm{Prk}\ \mathrm{Rek} / (\mathrm{Prk} + \mathrm{Rek}) \quad (18)$$

For a metric Mk, the macro and weighted averages are

$$\mathrm{Mmacro} = (1/K)\ \Sigma_k\ \mathrm{Mk},\quad \mathrm{Mwtd} = \Sigma_k\ (n_k / N)\ \mathrm{Mk} \quad (19)$$

Macro-averaged recall equals balanced accuracy. Because failing to flag a melanoma is the costliest error, the recall of MEL (and of BCC and AKIEC) should be reported explicitly.

C. Confusion Matrix and Class-Wise Performance

The confusion matrix has entries

$$\mathrm{CMij} = |\ \{n: y_n = i,\ \hat{y}_n = j\}\ | \quad (20)$$

so that the diagonal contains correct predictions, and the off-diagonal entries show which classes are confused. Class-wise precision, recall and F1-score are derived from it.

D. ROC-AUC

For each class, a receiver operating characteristic (ROC) curve plots recall against the false-positive rate (1 − specificity) for varying thresholds on P (yk | x) in a one-versus-rest setting. The area under the curve (AUC) is threshold-independent and is summarised by the macro average

$$\mathrm{AUCmacro} = (1/K)\ \Sigma_k\ \mathrm{AUCk} \quad (21)$$

AUC should be read alongside precision and recall, since in imbalanced data it can look favourable even when the minority-class precision is low.

E. Confidence Calibration

A model is calibrated when its confidence matches its accuracy. Following Guo et al. [19], the test predictions are grouped into B confidence bins Bb, and the expected calibration error is

$$ECE = \Sigma b \, (|Bb| / N) \, | \, acc \, (Bb) - conf \, (Bb) \, | \qquad (22)$$

where acc (Bb) is the accuracy and conf (Bb) the mean confidence in bin b. Reliability diagrams and, if necessary, temperature scaling is used to examine and reduce miscalibration. Until this analysis has been done, the confidence score must not be read as a probability of correctness.

F. Inference Time and Agent-Level Safety Measure

Mean and 95th-percentile inference time are measured for A2 alone and for the whole pipeline (A1–A7) on the target hardware, using the same images and batch size as in deployment. For the decision layer, the referral sensitivity measures how often a truly concerning lesion is routed to professional evaluation:

$$Seref = Nref / NG \qquad (23)$$

where NG is the number of test images whose true class is in G = {MEL, BCC, AKIEC}, and Nref the number of those for which A4 assigns a risk level other than LOWER CONCERN. It should be reported together with the referral rate for benign cases, because over-referral reduces the usefulness of the system.

Table V. Evaluation Framework

Measure	Level	Purpose
Accuracy, balanced accuracy	Overall	Headline performance
Precision, recall, F1	Per class, macro, weighted	Class-wise reliability
Confusion matrix	Per class pair	Error structure
ROC-AUC (one-vs-rest)	Per class, macro	Ranking quality
ECE, reliability diagram	Confidence bins	Calibration
Inference time	A2 and A1–A7	Responsiveness
Referral sensitivity Seref	Classes in G	Safety of A4

X. RESULTS AND DISCUSSION

This section is the template for reporting the measured results. Entries in square brackets are placeholders, to be filled in with values from the implementation. No experimental result is reported or implied in the current draft, and no claim of superiority over any existing system is made.

A. Dataset Distribution

The class distribution of the full dataset is given in Table III. The numbers of training, validation and test images after the lesion-level split are: training [Insert number], validation [Insert number], test [Insert number].

B. Training and Validation Performance

[Insert figure: training and validation loss versus epoch.] [Insert figure: training and validation macro-F1 versus epoch.] The epoch at which early stopping occurred, the learning rates used in the two training phases and the training time are to be stated here: [Insert].

C. Classification Results

Table VI gives the overall test-set results and Table VII the class-wise results.

Table VI. Overall Test-Set Performance (to be measured)

Measure	Value
Accuracy	[Insert measured accuracy]
Balanced accuracy (macro recall)	[Insert value]
Macro precision / recall / F1	[Insert] / [Insert] / [Insert]
Weighted precision / recall / F1	[Insert] / [Insert] / [Insert]
Macro ROC-AUC	[Insert value]
ECE (before / after calibration)	[Insert] / [Insert]
Mean inference time: A2 / A1–A7	[Insert ms] / [Insert ms]
Referral sensitivity Seref	[Insert value]

Table VII. Class-Wise Test-Set Performance (to be measured)

Class	Prec.	Recall	F1	AUC	Support
MEL	[ ]	[ ]	[ ]	[ ]	[ ]
NV	[ ]	[ ]	[ ]	[ ]	[ ]
BCC	[ ]	[ ]	[ ]	[ ]	[ ]
AKIEC	[ ]	[ ]	[ ]	[ ]	[ ]
BKL	[ ]	[ ]	[ ]	[ ]	[ ]
DF	[ ]	[ ]	[ ]	[ ]	[ ]
VASC	[ ]	[ ]	[ ]	[ ]	[ ]

D. Confusion Matrix Analysis

Table VIII. Confusion Matrix (rows: true; columns: predicted; to be measured)

	MEL	NV	BCC	AKIEC	BKL	DF	VASC
MEL	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
NV	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
BCC	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
AKIEC	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
BKL	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
DF	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
VASC	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

The analysis should identify the largest off-diagonal entries and discuss them. Pairs that are visually similar, such as melanoma and melanocytic nevi, or benign keratosis-like lesions and nevi, are expected to be of interest, but whether they are the main sources of error must be established from the data: [Insert analysis]. The number of true MEL, BCC and AKIEC cases assigned to a benign class with a high C is of particular importance and should be reported.

E. Agent Workflow Outputs

Table IX illustrates how A3–A5 combine to produce risk levels. These are hypothetical examples that demonstrate the rule logic of Table IV; they are not measured outputs.

Table IX. Illustrative Rule Logic (hypothetical cases)

#	A2/A3 output	A4 rule and level	Recommendation
1	NV, $C \geq T_{conf}$, $S_{conc} < T_{conc}$	R5: LOWER CONCERN	Monitor; seek evaluation if it changes
2	BKL, $C < T_{conf}$	R2: UNCERTAIN	Professional evaluation
3	NV, $C \geq T_{conf}$, $S_{conc} \geq T_{conc}$	R4: ELEVATED	Professional evaluation
4	MEL, $C \geq T_{conf}$	R3: HIGH CONCERN	Prompt professional evaluation
5	Image fails quality check	R1: RECAPTURE	Upload a clearer image

For the implemented system, representative end-to-end reports for confident, uncertain and high-concern cases are to be inserted: [Insert example reports].

F. Limitations and Observations

Accuracy values reported in the literature, such as those cited in Section III, were obtained with different datasets, subsets, augmentation and splitting strategies. They cannot be compared directly with results from this system. In particular, an evaluation in which several images of one lesion are present in both the training and the test sets overestimates performance. The comparison with the literature therefore has to be made only after the protocol of Section VI has been followed. Observations on failure cases, uncertain cases and the effect of the thresholds: [Insert].

XI. ADVANTAGES AND APPLICATIONS

Automated preliminary image assessment. The system provides a fast, consistent first look at a lesion image for the seven HAM10000 categories, which may help to prioritise cases.

Modular multi-agent architecture. Each agent has one responsibility and a typed interface. The classifier, the rules or the report template can be changed independently, and each agent can be tested separately.

Confidence-aware output. The system reports the probability vector and confidence indicators, and treats low confidence as a reason for escalation instead of presenting an uncertain label as a finding.

Structured assessment reports. A fixed report format records the prediction, indicators, risk level, rationale and guidance, which supports review, auditing and communication between a patient and a clinician.

General skin-care guidance. Non-prescriptive guidance on sun protection, skin hygiene and self-monitoring is supplied with each report.

Teledermatology support. In remote consultations, the risk level and uncertainty flag could help triage incoming images, in line with the observation that AI support can help clinicians when it is well presented [16].

Healthcare decision-support applications. The architecture can serve as a pattern for other image-based screening tasks in which a classifier is followed by reliability checks, rules and reporting.

Future integration. Additional datasets, models and explanation modules can be added by extending or replacing individual agents.

XII. LIMITATIONS AND FUTURE SCOPE

A. Limitations

Dataset. HAM10000 contains seven categories of pigmented lesions from two sources and is strongly imbalanced [4]. The system cannot recognise other skin conditions, and for such images it may still produce one of the seven classes. Input images that are not dermatoscopic, for example smartphone photographs, may differ strongly from the training distribution.

Image quality. Poor focus, lighting, hair, ruler marks or ink can degrade the prediction. A1 performs only simple quality checks.

Skin-tone representation and bias. Reviews of dermatology AI datasets report incomplete description of patient diversity and skin tone [18]. A model trained on HAM10000 should not be assumed to perform equally across skin tones, ages or body sites, and subgroup evaluation is needed.

Generalisation. Algorithms that perform well on a test set may generalise less well to images from sources that were not represented in training [9]. External validation is therefore essential.

Confidence. Softmax confidence is not a calibrated probability, and deep networks tend to be over-confident [19]. The thresholds of A3 and A4 are design parameters that are not clinically validated.

Clinical validation. No clinical validation is claimed. Prospective studies with dermatologists, including reader studies of the kind described in [16], are required before any clinical use. The system must not be used as a diagnostic tool or as a substitute for professional evaluation.

Agent design. The agents are deterministic modules that follow a fixed pipeline. They cannot adapt their strategy, negotiate or learn from feedback.

B. Future Scope

Future work includes training on larger and more diverse, multi-site datasets with documented skin-tone information, and adding further dermatological categories, including non-pigmented lesions. Explainability can be added by integrating gradient-based saliency such as Grad-CAM [12], so that the report displays the image regions behind the prediction. Uncertainty estimation can be improved with temperature scaling, Monte-Carlo dropout, deep ensembles or conformal prediction. The system can be packaged as a web and mobile application, connected to professional healthcare and teledermatology workflows, and evaluated in prospective studies. Where justified, an LLM-based agent could be added to phrase reports or explanations, subject to safety checks, and the coordination layer could be extended following the taxonomy of [7].

XIII. CONCLUSION

This paper addressed the problem that skin-lesion classifiers usually stop at a class label, although a clinical setting needs to know how far a prediction can be relied on, what action it implies and how it is communicated. We proposed an Agentic AI-based dermatological assessment and skin-care decision system in which seven specialized agents handle preprocessing, classification, confidence evaluation, risk assessment, guidance, coordination and reporting. Classification is performed by a single EfficientNet-B0 network fine-tuned by transfer learning for the seven HAM10000 categories, which keeps the model compact. The coordination agent sequences the others through a shared case record, so that all decisions can be traced. Risk assessment follows transparent predefined rules that escalate uncertain or potentially concerning results to professional evaluation.

The expected contribution is a design and an evaluation framework, not a performance claim: a documented separation of learned and rule-based components, algorithms with explicit mathematical definitions, and an evaluation protocol designed for class imbalance, lesion-level splitting and calibration. The work has clear limitations. The classifier covers only seven pigmented-lesion categories, confidence is uncalibrated, the thresholds are not clinically validated, dataset bias and skin-tone representation are open issues, and the report sections contain placeholders for results that have not yet been measured. Future work will complete the experiments, calibrate and validate the decision layer, add explainability and stronger uncertainty estimation, extend the categories and data, and evaluate the system with dermatologists. Throughout, the system is a preliminary decision-support aid, and professional dermatological evaluation remains necessary for any medical decision.

ACKNOWLEDGMENT

The authors thank [Guide / Institution name] for guidance and support. The HAM10000 dataset was made available by its creators for research use [4].

REFERENCES

[1]. H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," Lancet Oncol., vol. 3, no. 3, pp. 159–165, Mar. 2002, doi: 10.1016/S1470-2045(02)00679-4.

[2]. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[3]. H. A. Haenssle et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," Ann. Oncol., vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[4]. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-sources dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, Art. no. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

[5]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[6]. M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," Knowl. Eng. Rev., vol. 10, no. 2, pp. 115–152, Jun. 1995, doi: 10.1017/S0269888900008122.

[7]. A. L. Silveira, R. da Rosa Righi, and C. A. da Costa, "Multi-agent systems for clinical decision support: A systematic review," Appl. Soft Comput., vol. 188, Art. no. 114447, Feb. 2026, doi: 10.1016/j.asoc.2025.114447.

[8]. M. E. Celebi et al., "A methodological approach to the classification of dermoscopy images," Comput. Med. Imaging Graph., vol. 31, no. 6, pp. 362–373, Sep. 2007, doi: 10.1016/j.compmedimag.2007.01.003.

[9]. P. Tschandl et al., "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study," Lancet Oncol., vol. 20, no. 7, pp. 938–947, Jul. 2019, doi: 10.1016/S1470-2045(19)30333-X.

[10]. S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[11]. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[12]. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[13]. G. Munjal, P. Bhardwaj, V. Bhargava, S. Singh, and N. Nagpal, "SkinSage XAI: An explainable deep learning solution for skin lesion diagnosis," Health Care Sci., vol. 3, no. 6, pp. 438–455, Dec. 2024, doi: 10.1002/hcs2.121.

[14]. D. Chanda, M. S. H. Onim, H. Nyeem, T. B. Ovi, and S. S. Naba, "DCENSnet: A new deep convolutional ensemble network for skin cancer classification," Biomed. Signal Process. Control, vol. 89, Art. no. 105757, Mar. 2024, doi: 10.1016/j.bspc.2023.105757.

[15]. H. A. Owida et al., "A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic images," Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-21783-z.

[16]. P. Tschandl et al., "Human–computer collaboration for skin cancer recognition," Nat. Med., vol. 26, no. 8, pp. 1229–1234, Aug. 2020, doi: 10.1038/s41591-020-0942-0.

[17]. N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.

[18]. R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review," JAMA

Dermatol., vol. 157, no. 11, pp. 1362–1369, Nov. 2021, doi: 10.1001/jamadermatol.2021.3129.

[19]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), Sydney, NSW, Australia, 2017, pp. 1321–1330.

[20]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, arXiv:1412.6980.