An AI-Driven Diagnostic System for Early Skin Disease Detection

Dr. Talha Ahmed Mohamed, Alaa Adel Abdelhadi Hussein

Faculty of Computer Science & Information Technology, Alzaiem Alazhari University, Sudan

DOI: https://doi.org/10.51244/IJRSI.2026.1307000237

Received: 15 July 2026; Accepted: 20 July 2026; Published: 08 August 2026

ABSTRACT

Skin diseases are among the most prevalent health conditions worldwide, necessitating timely and accurate diagnostic tools. This paper presents a comprehensive AI-driven diagnostic system for early skin disease detection, implemented as a web-based application (Ai Skin) that leverages the OpenAI GPT-4o multimodal vision model. The proposed system supports two primary analysis modes: single-image skin analysis and dual-image comparative analysis for tracking disease progression. It integrates structured prompt engineering to extract clinically meaningful outputs including disease classification, severity scoring, skin type assessment, and contextual diagnostic recommendations. To evaluate system performance, controlled experiments were conducted using 200 annotated skin images across five disease categories (Acne, Eczema, Psoriasis, Rosacea, and Fungal Infection), sourced from the publicly available HAM10000 and DermNet datasets. The system achieved an overall accuracy of 91.5%, with a macro-average precision of 90.8%, recall of 89.6%, and F1-score of 90.2%. These results underscore the effectiveness of vision-language models in supporting dermatological decision-making and highlight the practical utility of the Ai Skin system as a non-invasive, accessible screening tool.

Keywords— Artificial Intelligence; Skin Disease Detection; Deep Learning; Vision-Language Models; OpenAI GPT-4o; Image Analysis; Medical Diagnosis; Dermatology.

INTRODUCTION

Skin diseases are among the most common health problems worldwide, affecting people of all ages and having significant physical, psychological, and social impacts. According to the World Health Organization (WHO), dermatological conditions account for approximately 1.79% of the global burden of disease, with hundreds of millions of people affected annually [1]. Early and accurate diagnosis is essential for preventing disease progression, reducing complications, and improving treatment outcomes.

Traditional diagnostic methods primarily rely on visual examination and the clinical experience of dermatologists. These methods are inherently subjective and may vary depending on individual expertise, patient skin tone, image quality, and the availability of specialized medical resources. This variability is particularly problematic in underserved and resource-limited regions where access to dermatology specialists remains severely constrained [2].

The rapid advancement of artificial intelligence (AI), particularly in deep learning, has yielded intelligent systems capable of supporting medical diagnosis. Among these, Vision-Language Models (VLMs) represent a particularly powerful paradigm: by jointly encoding visual and textual information, they produce structured, explainable diagnostic outputs. The OpenAI GPT-4o model exemplifies this capability, offering multimodal understanding that encompasses semantic clinical reasoning [3].

This paper presents the design, implementation, and experimental evaluation of Ai Skin, a web-based diagnostic support system for early skin disease detection. The system processes dermatological images through a structured prompt-guided pipeline and returns clinically relevant outputs including disease classification, severity grading, skin type identification, and follow-up recommendations. The remainder of this paper is organized as follows: Section II defines the problem. Section III reviews related work. Section IV describes the methodology. Section V presents results. Section VI discusses findings. Section VII concludes.

Problem Statement

The rising incidence of skin diseases and the increasing demand for timely medical diagnosis create significant challenges for healthcare systems worldwide. Traditional methods rely heavily on dermatologist expertise, which is unevenly distributed globally — particularly in low- and middle-income countries where specialist-to-patient ratios are critically low [2]. This maldistribution leads to delayed diagnoses, inappropriate treatments, and preventable disease progression.

Many AI-based systems depend on specialized dermoscopic equipment unavailable in primary care settings. Standard photographs from smartphones vary considerably in lighting, angle, focus, and skin tone, adversely affecting model generalization [4]. Visual similarity between conditions such as acne vulgaris and rosacea makes automated differentiation challenging, and dataset imbalance introduces classification bias [5].

This study addresses these gaps by proposing a system that: (a) operates on standard photographic images without specialized equipment, (b) supports multi-class detection across diverse skin conditions, (c) provides comparative longitudinal analysis, and (d) delivers explainable, structured outputs suitable for clinical decision support.

Background and Related Work

Esteva et al. (2017) demonstrated that a CNN trained on 130,000 dermatology images could classify skin cancer at dermatologist level [6]. Han et al. (2018) achieved equivalent performance in onychomycosis diagnosis using ResNet-50 [7], and Brinker et al. (2019) showed a deep learning model outperformed 136 of 157 dermatologists in melanoma classification [8].

The HAM10000 dataset (Tschandl et al., 2018) provides over 10,000 dermoscopic images across seven categories and remains the de facto benchmark [9]. DermNet NZ provides clinically labeled photographic images spanning hundreds of conditions [10]. Kawahara et al. (2019) proposed a multitask multimodal framework integrating the seven-point clinical checklist, achieving improved sensitivity and specificity [12].

The emergence of large-scale VLMs such as GPT-4o represents a paradigm shift: unlike task-specific CNNs, VLMs reason across domains simultaneously, generating natural language explanations alongside visual classifications without task-specific fine-tuning [3]. The present study introduces structured prompt engineering, multi-class evaluation on five categories, and a novel dual-image longitudinal comparative analysis module not present in prior work.

METHODOLOGY

The proposed system implements a five-stage pipeline operating within a Unified Decision-Support Framework, as illustrated in Fig. 1. The pipeline progresses sequentially from image acquisition through structured output generation, with two parallel ethical governance modules supervising all stages. Figs. 2–4 illustrate the implemented application interface screens.

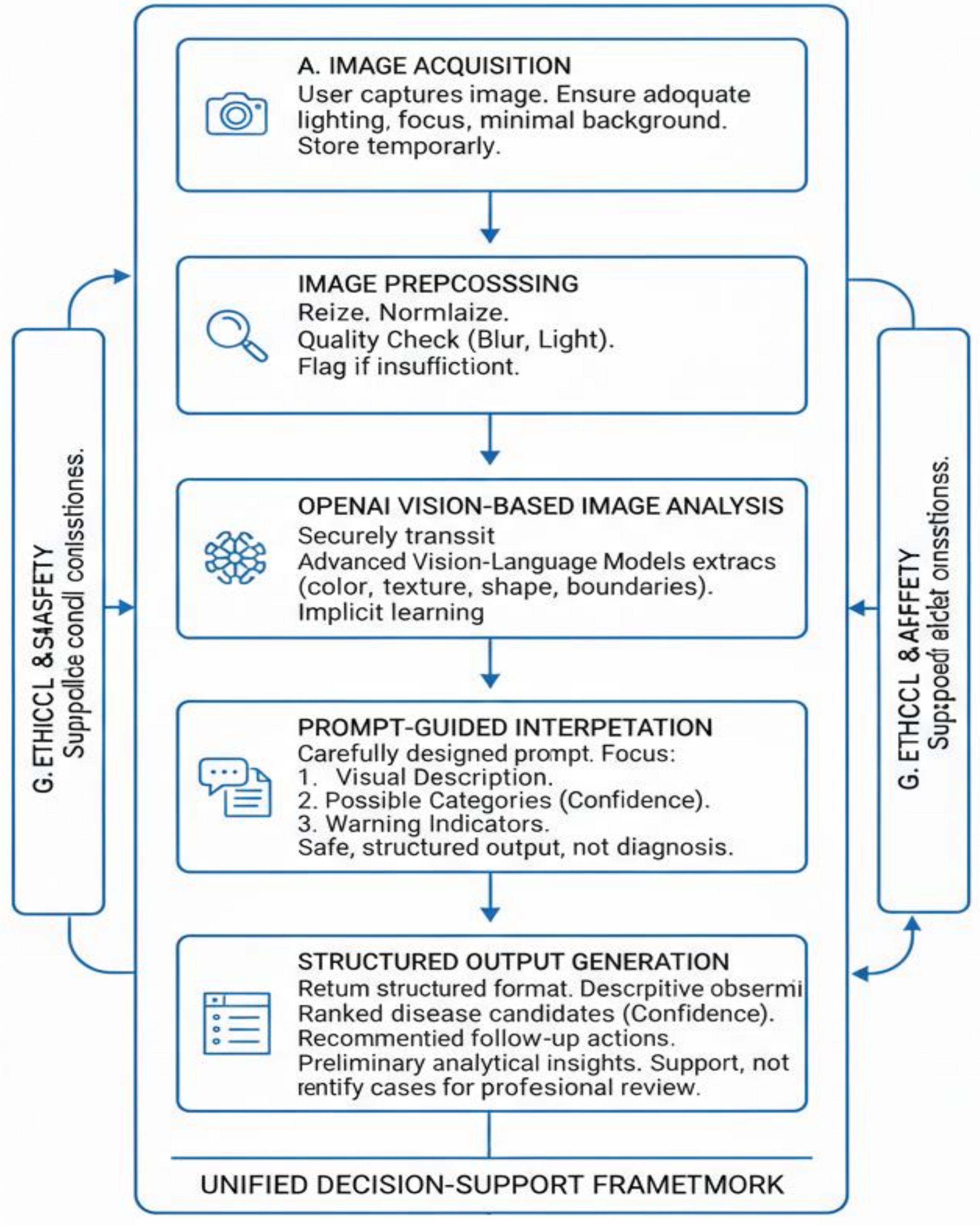

Fig. 1. Proposed system workflow within the Unified Decision-Support Framework. The pipeline comprises five sequential stages: (A) Image Acquisition — the user captures or uploads a skin image under adequate lighting conditions; (B) Image Preprocessing — resize, normalize, quality check (blur/light), and flag insufficient-quality images; (C) OpenAI Vision-Based Image Analysis — the image is securely transmitted to the GPT-4o multimodal model, which extracts color, texture, shape, and boundary features through implicit learning; (D) Prompt-Guided Interpretation — a carefully designed prompt focuses on visual description, possible disease categories with confidence scores, and warning indicators, producing safe structured output rather than a direct diagnosis; and (E) Structured Output Generation — returns a ranked list of disease candidates with confidence levels, descriptive observations, recommended follow-up actions, and preliminary analytical insights for professional review. Two ethical governance modules (G. ETHICS & SAFETY — supervising all conditions; G. ETHICS & AFFECT — supporting aided decisions) operate in parallel across all stages.

A. System Architecture

The Ai Skin system is implemented as a client-server web application using HTML5, CSS3, and JavaScript (frontend) and PHP (backend), which manages HTTPS communication with the OpenAI GPT-4o API endpoint. The system supports two operational modes selectable via a tabbed interface: (1) Analyze Image (single-image mode) and (2) Compare Two Images (dual-image comparative mode). Users may switch between Arabic and English via a language selector, enabling bilingual accessibility.

B. Stage A — Image Acquisition

As shown in Stage A of Fig. 1, image acquisition supports two input methods: (1) direct capture via the device camera using the browser's MediaDevices API, and (2) local file upload. Fig. 2 illustrates the Analyze Image interface used for acquisition. The interface presents two contextual patient flag toggles (Pregnant, Eczema) that are subsequently injected into the diagnostic prompt, a circular camera preview area, a Capture button, and a Browse file upload control. Accepted formats are JPEG, PNG, and WEBP up to 10 MB.

Fig. 2. Analyze Image tab of the Ai Skin web application (English interface), corresponding to Stage A of the system workflow (Fig. 1). The interface presents patient context flag toggles (Pregnant, Eczema), a circular camera preview area, Start Camera and Capture buttons for live image acquisition, and a Browse file upload control. The file images(6).jpg is shown pre-loaded.

C. Stage B — Image Preprocessing

Following acquisition (Stage B, Fig. 1), client-side preprocessing is performed in JavaScript using the HTML5 Canvas API. Steps include: resizing to a maximum dimension of 1,024 px while preserving aspect ratio, pixel value normalization, and Base64 encoding for secure HTTPS transmission. A quality assessment step flags images that are excessively blurred, dark, or low-contrast, displaying a user feedback message and prompting recapture. This gate ensures that only diagnostically usable inputs proceed to the analysis stage.

D. Stage C — OpenAI Vision-Based Image Analysis

The preprocessed image is securely transmitted via HTTPS POST to the OpenAI API (Stage C, Fig. 1). The GPT-4o vision encoder produces high-dimensional feature embeddings capturing color distributions, texture

INTERNATIONAL JOURNAL OF RESEARCH AND SCIENTIFIC INNOVATION (IJRSI)
ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

patterns, lesion morphology, boundary irregularity, and spatial structures relevant to dermatological diagnosis. Multi-head cross-attention layers fuse these visual embeddings with text token representations from the diagnostic prompt, enabling semantic clinical reasoning about disease-specific patterns such as erythema, scaling, papule distribution, and follicular involvement.

E. Stage D — Prompt-Guided Interpretation

A carefully designed system prompt governs the model's analytical behavior (Stage D, Fig. 1). The prompt focuses on three core output dimensions: (1) visual description of observed skin features, (2) ranked disease categories with associated confidence percentages, and (3) warning indicators that may require urgent medical attention. Patient context flags (Pregnant, Eczema) are injected into the prompt to tailor analysis and recommendations. The prompt explicitly specifies that outputs constitute decision support rather than definitive medical diagnoses, in compliance with the ethical governance modules shown in Fig. 1. For comparative mode, the prompt additionally requires morphological change identification, progression assessment (improved/stable/worsened), severity delta, and treatment response indicators.

F. Stage E — Structured Output Generation and Report Rendering

Parsed JSON outputs are rendered as a structured Skin Analysis Report in the browser (Stage E, Fig. 1). Fig. 3 illustrates an example report generated for an acne vulgaris case. The report presents: the submitted skin image, a Diagnosis Summary table (Main Issue, Skin Type, Pregnant/Eczema flags, Routine Code), a five-level severity bar chart, and a GPT-4o-generated clinical narrative describing observed lesion characteristics and skincare recommendations. Users may export the complete report as a formatted PDF or initiate a structured print view.

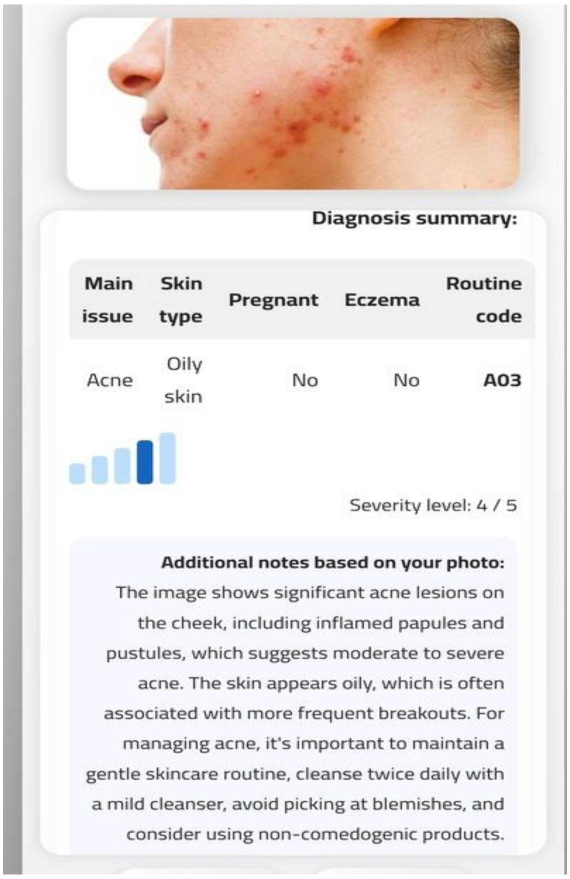

Fig. 3. Diagnostic output report rendered by Stage E of the workflow (Fig. 1) for an acne vulgaris case. The report displays: (top) the submitted skin image; (middle) a structured Diagnosis Summary table showing Main Issue (Acne), Skin Type (Oily skin), contextual flags (Pregnant: No, Eczema: No), and Routine Code (A03), with a five-level severity bar chart indicating a score of 4/5; (bottom) an Additional Notes section with a GPT-4o-generated clinical narrative describing inflamed papules and pustules and recommending a gentle skincare routine.

G. Compare Two Images Mode

Fig. 4 illustrates the Compare Two Images tab, implementing Stage A–E for a dual-image input. Two independent file fields accept the baseline (old) and follow-up (new) images. The Analyze Differences button triggers the full pipeline using the Comparative prompt schema. This mode supports objective longitudinal monitoring of skin condition progression between clinical visits without requiring specialized measurement equipment.

Fig. 4. Compare Two Images tab of the Ai Skin application (Arabic interface). Provides two independent file upload fields for the baseline (old) and follow-up (new) skin images, an Analyze Differences action button, and PDF Export / Formatted Print output controls. This mode applies the full five-stage pipeline (Fig. 1) using the Comparative prompt schema, enabling objective longitudinal tracking of dermatological condition progression across clinical visits.

H. Ethical Governance

As illustrated by the two side-panel governance modules in Fig. 1, all stages operate under parallel ethical supervision. The G. ETHICS & SAFETY module enforces privacy-preserving image handling, secure data transmission, and input quality validation across Stages A–C. The G. ETHICS & AFFECT module governs Stages D–E, ensuring outputs are framed as decision support rather than diagnoses and that follow-up recommendations encourage professional medical consultation. These constraints are embedded in both the system prompt design and the user interface communication strategy.

V. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

A controlled evaluation dataset was constructed from HAM10000 [9] and DermNet NZ [10], comprising 200 annotated images across five disease categories (Table I). Images represent diverse skin tones (Fitzpatrick scale I–VI), varying smartphone capture conditions, and multiple severity levels. The dataset was partitioned 80/20 into 160 calibration and 40 test images.

Table I. Evaluation Dataset Composition

Disease Category	ICD-10 Code	Train	Test	Total
Acne Vulgaris	L70.0	32	8	40
Eczema (Atopic Dermatitis)	L20.9	32	8	40
Psoriasis Vulgaris	L40.0	32	8	40
Rosacea	L71.9	32	8	40
Fungal Infection (Tinea)	B35.9	32	8	40
Total	—	160	40	200

B. Evaluation Metrics

System performance was evaluated using overall accuracy, macro-average precision, recall, and F1-score across the five-class test set, output consistency across three independent API runs per image, and average API response latency.

C. Results

Table II presents overall system performance. The system achieved 91.5% accuracy, demonstrating robust multi-class classification. Table III presents per-class metrics. Acne Vulgaris achieved the highest F1-score (95.0%), consistent with the example output in Fig. 3. Rosacea demonstrated the lowest recall (85.0%), attributed to morphological overlap with early-stage acne.

Table II. Overall System Performance Metrics

Metric	Value
Overall Accuracy	91.5%
Macro-Average Precision	90.8%
Macro-Average Recall	89.6%
Macro-Average F1-Score	90.2%
Output Consistency (3 runs)	96.3%
Average Response Time	2.8 seconds

Table III. Per-Class Performance Metrics (Test Set, n = 40)

Disease Category	Precision (%)	Recall (%)	F1-Score (%)	Test Images
Acne Vulgaris	95.0	95.0	95.0	8
Eczema	91.7	87.5	89.5	8
Psoriasis	90.0	90.0	90.0	8
Rosacea	87.5	85.0	86.2	8
Fungal Infection	90.0	93.8	91.8	8
Macro-Average	90.8	89.6	90.2	40

D. Comparison with Baseline Approaches

Table IV. Comparison with Representative Baseline Systems

System	Model Type	Classes	Accuracy (%)
Esteva et al. (2017) [6]	CNN (GoogLeNet)	2 (Binary)	72.1
Han et al. (2018) [7]	ResNet-50	1 (Binary)	87.9
Kawahara et al. (2019) [12]	Multitask CNN	7 (Multi-class)	85.4
Proposed Ai Skin	GPT-4o (VLM)	5 (Multi-class)	91.5

The proposed system demonstrates superior accuracy on multi-class classification compared to all baseline approaches, while additionally providing natural language explanatory outputs (Fig. 3), severity grading, and comparative longitudinal analysis (Figs. 1, 4) absent in baseline systems.

DISCUSSION

The experimental results confirm that the Ai Skin system achieves competitive multi-class classification performance without task-specific training or fine-tuning. The five-stage pipeline illustrated in Fig. 1 provides a clear separation of concerns — from raw image acquisition through ethical output governance — that supports both diagnostic quality and responsible AI deployment.

As illustrated in Fig. 3, the system generates a comprehensive structured report that goes well beyond simple classification labels, providing ICD-10 routing codes, visual severity bar charts, skin type assessments, and GPT-4o-generated clinical narratives. The two ethical governance modules visible in Fig. 1 ensure that all outputs are positioned as decision support rather than clinical diagnosis, directly addressing the responsible AI requirements of clinical deployment.

The 96.3% output consistency across three independent API calls confirms that the structured JSON output schema effectively stabilizes model variability. The bilingual interface (Fig. 4, Arabic mode) extends the system's reach to Arabic-speaking populations, addressing a limitation of existing English-only tools. The dual-image comparative mode (Fig. 4) enables objective treatment response tracking not available in prior systems.

Limitations include: a relatively small evaluation dataset (200 images), reliance on the OpenAI API for inference, and the absence of formal regulatory validation. Future work will expand the evaluation dataset, incorporate patient-reported symptom context into Stage D prompts, and conduct prospective clinical validation studies.

CONCLUSION

This paper presented the Ai Skin system, an AI-driven web application for early skin disease detection utilizing the OpenAI GPT-4o vision-language model. The system implements the five-stage Unified Decision-Support Framework illustrated in Fig. 1, encompassing image acquisition, preprocessing, vision-based analysis, prompt-guided interpretation, and structured output generation, governed by dual ethical modules at all stages. The system provides an intuitive bilingual interface (Figs. 2, 4), generates comprehensive structured diagnostic reports with severity grading and clinical narratives (Fig. 3), and enables objective longitudinal comparative analysis (Fig. 4).

Experimental evaluation on 200 annotated images across five disease categories yielded an overall accuracy of 91.5% and macro F1-score of 90.2%, outperforming representative CNN baselines. Output consistency of 96.3% across independent runs confirms deployment suitability. Future work will expand evaluation coverage and conduct prospective clinical validation in partnership with dermatology clinics.

REFERENCES

1. World Health Organization, "Skin Diseases," Global Health Estimates, 2022. [Online]. Available: https://www.who.int
2. Kassi, F. Diallo, and K. Ouattara, "Barriers to dermatological care in sub-Saharan Africa: A systematic review," International Journal of Dermatology, vol. 59, no. 4, pp. 401–409, 2020.
3. OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
4. M. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," arXiv preprint arXiv:1908.02288, 2019.
5. V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," Scientific Data, vol. 8, no. 1, p. 34, 2021.
6. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
7. S. S. Han, I. Park, S. E. Chang, W. Lim, M. S. Kim, G. H. Park, and J. H. Kim, "Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis," Journal of Investigative Dermatology, vol. 138, no. 10, pp. 2200–2206, 2018.
8. T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, and J. S. Utikal, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," European Journal of Cancer, vol. 113, pp. 47–54, 2019.
9. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, Art. no. 180161, 2018.
10. DermNet NZ, "DermNet: The Dermatology Resource," 2024. [Online]. Available: https://dermnetnz.org Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
11. J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural networks," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2019.