Master Thesis

Real-World Evaluation of GPT-4 Turbo Vision for Dermoscopic Skin Cancer Detection

submitted by
Paula Ďuríková, MUDr.

in partial fulfillment of the requirements for the degree of

Master of Science (Continuing Education) Msc (CE)

at the

Medical University of Graz

executed in the
Master of Dermoscopy and Preventive Dermatooncology

under the supervision of Harald Kittler, Ao.Univ.-Prof. Dr.med.univ.

Bratislava 30.7.2025

I hereby declare that the submitted thesis is my own work. All direct and indirect sources are acknowledged in the references. This Master's thesis has not been used in the same or similar version to obtain any academic degree.

Paula Ďuríková Bratislava 17.9.2025

Table of Contents

Title:

Real-World Evaluation of GPT-4 Turbo Vision for Dermoscopic Skin Cancer Detection

Abstract

Background: Generative artificial intelligence (AI) models, such as ChatGPT, are increasingly being integrated into dermatology for skin cancer detection, despite limited validation. This study simulates real-world primary care workflows to bridge the gap between AI innovation and clinical application. The findings aim to determine whether integrating generative AI into dermatologic triage is feasible, particularly in underserved settings with limited specialist access. Additionally, diagnostic uncertainty metrics identified areas that require human-AI collaboration. This protocol adheres to the latest Standards for Reporting Diagnostic Accuracy Studies (STARD 2015) and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Artificial Intelligence Extension (TRIPOD-AI) guidelines.

Methods and analysis: In this retrospective, cross-sectional diagnostic accuracy study, 750 prospectively curated dermoscopic images (533 benign, 217 malignant/precancerous) from validated repositories (International Skin Imaging Collaboration Archive, DermNet New Zealand, dermatology textbooks) were analyzed. Lesions spanned melanocytic, non-melanocytic, keratinocyte-derived, and vascular tumors, reflecting real-world diagnostic diversity. The latest publicly available version of GPT-4 Turbo with Vision (OpenAI) provided diagnoses via standardized, clinician-mimicking prompts without clinical metadata. The reference standard consists of histopathologically or expert-confirmed diagnoses. Primary outcome: overall diagnostic accuracy. Secondary outcomes: diagnostic certainty (re-prompting frequency, unclassifiable responses), lesion-specific sensitivity/specificity, triage performance (benign, precancerous, malignant), explainability, and management suggestions.

Results: GPT-4 Turbo with Vision demonstrated an overall diagnostic accuracy of 37.6% (282/750; 95% CI 34.1–41.1) across eight lesion categories. Sensitivity varied widely: highest for melanoma (0.64, 95% CI 0.50–0.76) and lowest for actinic keratosis (0.07, 95% CI 0.02–0.18). Broad specificity ranged from 0.85 (nevi) to 0.98 (squamous cell carcinoma/actinic keratosis). In mimic-specific comparisons, specificity was 0.80 for melanoma vs. atypical

nevi and 0.86 for squamous cell carcinoma vs. actinic keratosis. Cohen's kappa showed moderate agreement for nevi (κ=0.45) and basal cell carcinoma (κ=0.36) but minimal concordance for squamous cell carcinoma/actinic keratosis ($\kappa \leq 0.17$). The model required follow-up reprompting in 32 cases (4.3%, 95% CI 3.0–5.9), with no unclassifiable images (0·0%, 95% CI 0·0–0·5). For triage classification (benign/precancerous/malignant), accuracy was 48.8% (366/750; 95% CI 45.2–52.4), with malignant lesions showing the highest sensitivity (0.77, 95% CI 0.70–0.83) and precancerous lesions the lowest (0.11, 95% CI 0.07–0.18). Management recommendations aligned with reference standards in 27.7% (208/750; 95% CI 24.8–30.9) of cases. Explainability analysis revealed a mean of 3.17 correctly identified dermoscopic features per lesion (95% CI 3.08–3.25), with melanocytic nevi (3.82) and melanoma (3.76) achieving the highest feature-level accuracy.

Conclusions: This study highlights significant limitations in the diagnostic performance of GPT-4 Turbo with Vision for dermoscopic image analysis, with an overall accuracy of 37.6% across benign, precancerous, and malignant lesions. While the model shows some ability to detect melanoma (sensitivity 0.64), its low sensitivity for actinic keratosis (0.07) and inconsistent triage accuracy (48.8%) underscore key areas for improvement. The suboptimal performance likely reflects the model's lack of specialized training in dermoscopy, which relies on pattern recognition distinct from general image interpretation. These findings suggest that, in its current form, this AI tool is not yet reliable for clinical decision support. However, given the rapid evolution of AI capabilities, further refinements—particularly in domain-specific training—may enhance its utility. For now, cautious interpretation of these results is warranted, with an emphasis on continued validation in real-world settings before considering integration into clinical workflows. The path forward should balance optimism for AI's potential with a measured assessment of its present limitations.

Funding: None.

Ethical Approval: Utilized de-identified, publicly available data; exempt from institutional review.

Registration: DOI 10.17605/OSF.IO/HKMTV

Keywords: Artificial intelligence; dermatology; dermoscopy; diagnostic accuracy; skin cancer; triage.

Titel:

Evaluation der realen Anwendung von GPT-4 Turbo Vision zur dermatoskopischen Hautkrebsdiagnose

Zusammenfassung

Hintergrund: Generative Künstliche Intelligenz (KI) wie ChatGPT werden zunehmend in der Dermatologie zur Hautkrebsdiagnose eingesetzt – trotz begrenzter Validierung. Diese Studie simuliert reale Abläufe in der hausärztlichen Versorgung, um die Lücke zwischen KI-Innovation und klinischer Anwendung zu schließen. Ziel ist es zu prüfen, ob sich generative KI sinnvoll in die dermatologische Triage integrieren lässt, insbesondere in unterversorgten Regionen mit eingeschränktem Zugang zu Fachärzt:innen. Zudem identifizieren Metriken zur diagnostischen Unsicherheit Bereiche, die eine Mensch-KI-Zusammenarbeit erfordern. Das Protokoll orientiert sich an den aktuellen Standards zur Berichterstattung diagnostischer Genauigkeit (STARD 2015) sowie den TRIPOD-AI-Richtlinien für transparente Modelle zur individuellen Prognose oder Diagnose.

Methoden und Analyse: In dieser retrospektiven, querschnittlichen Studie zur diagnostischen Genauigkeit wurden 750 prospektiv kuratierte dermatoskopische Bilder (533 benigne, 217 maligne/präkanzeröse) aus validierten Quellen (International Skin Imaging Collaboration Archive, DermNet New Zealand, dermatologische Fachliteratur) analysiert. Die Läsionen umfassten melanozytäre, nicht-melanozytäre, keratinozytäre und vaskuläre Tumoren, um die diagnostische Vielfalt in der Praxis abzubilden. Diagnosen wurden von der öffentlich verfügbaren Version von GPT-4 Turbo mit Vision (OpenAI) anhand standardisierter, klinisch orientierter Prompts ohne klinische Metadaten generiert. Der Referenzstandard waren histopathologisch oder durch Expert:innen bestätigte Diagnosen.

Primärer Endpunkt: Gesamtdiagnostische Genauigkeit.

Sekundäre Endpunkte: diagnostische Sicherheit (Re-Prompting-Frekvenz, nicht klassifizierbare Antworten), Läsions-spezifische Sensitivität/Spezifität, Triage-Leistung (benigne, präkanzerös, maligne), Erklärbarkeit und Therapieempfehlungen.

Ergebnisse: GPT-4 Turbo mit Vision erreichte eine Gesamtdiagnosegenauigkeit von 37,6 % (282/750; 95%-KI 34,1–41,1) über acht Läsionskategorien hinweg. Die Sensitivität variierte stark: am höchsten für Melanome (0,64; 95%-KI 0,50–0,76), am niedrigsten für aktinische

Keratosen (0,07; 95%-KI 0,02–0,18). Die Spezifität lag breit gestreut zwischen 0,85 (Nävi) und 0,98 (Plattenepithelkarzinome/aktinische Keratose). In Differenzierungen mit ähnlichem Erscheinungsbild betrug die Spezifität 0,80 für Melanom vs. atypische Nävi und 0,86 für Plattenepithelkarzinom vs. aktinische Keratose. Der Cohen's Kappa zeigte moderate Übereinstimmung für Nävi (κ=0,45) und Basalzellkarzinome (κ=0,36), jedoch nur minimale Übereinstimmung für Plattenepithelkarzinome/aktinische Keratosen ($\kappa \leq 0{,}17$). In 32 Fällen (4,3 %, 95%-KI 3,0–5,9) war ein Re-Prompting nötig; es gab keine unklassifizierbaren Bilder (0,0 %, 95%-KI 0,0–0,5). Die Triage-Genauigkeit (benigne/präkanzerös/maligne) lag bei 48,8 % (366/750; 95%-KI 45,2–52,4), wobei maligne Läsionen die höchste Sensitivität (0,77; 95%-KI 0,70–0,83) und präkanzeröse die niedrigste (0,11; 95%-KI 0,07–0,18) zeigten. Therapieempfehlungen stimmten in 27,7 % (208/750; 95%-KI 24,8–30,9) mit den Referenzstandards überein. Die Analyse der Erklärbarkeit ergab durchschnittlich 3,17 korrekt identifizierte dermatoskopische Merkmale pro Läsion (95%-KI 3,08–3,25), mit den höchsten Werten bei melanozytären Nävi (3,82) und Melanomen (3,76).

Fazit: Diese Studie zeigt deutliche Einschränkungen der diagnostischen Leistungsfähigkeit von GPT-4 Turbo mit Vision bei der Analyse dermatoskopischer Bilder, mit einer Gesamtdiagnosegenauigkeit von nur 37,6 % über benigne, präkanzeröse und maligne Läsionen hinweg. Trotz gewisser Stärken bei der Erkennung von Melanomen (Sensitivität 0,64) offenbaren die niedrige Sensitivität bei aktinischer Keratose (0,07) und die uneinheitliche Triage-Leistung (48,8 %) zentrale Schwächen. Die mangelhafte Leistung dürfte auf das fehlende domänenspezifische Training im Bereich der Dermatoskopie zurückzuführen sein, die stark auf Mustererkennung basiert – ein Bereich, der sich deutlich von allgemeiner Bildinterpretation unterscheidet. Die Ergebnisse legen nahe, dass dieses KI-Tool in seiner aktuellen Form nicht für klinische Entscheidungen geeignet ist. Angesichts der rasanten Weiterentwicklung von KI könnten jedoch gezielte Weiterentwicklungen – insbesondere spezialisiertes Training – die Anwendbarkeit verbessern. Bis dahin sind die Ergebnisse mit Vorsicht zu interpretieren, und eine sorgfältige Validierung unter realen Bedingungen ist unerlässlich, bevor eine Integration in den klinischen Alltag in Betracht gezogen wird.

Förderung: Keine.

Ethik: Nutzung anonymisierter, öffentlich zugänglicher Daten; von der ethischen Prüfung

ausgenommen.

Registrierung: DOI 10.17605/OSF.IO/HKMTV

Schlüsselwörter: Künstliche Intelligenz; Dermatologie; Dermatoskopie; Diagnostische Genauigkeit; Hautkrebs; Triage.

1 Introduction

The role of generative artificial intelligence (AI) models, such as GPT-4 (Generative Pre-trained Transformer) Turbo with Vision, in dermoscopy is still in its early stages and has not yet been fully integrated into clinical practice. Despite the absence of comprehensive validation studies assessing their diagnostic accuracy, these models are increasingly integrated into dermatology due to their rapid processing and broad accessibility. Unlike specialized diagnostic software, which is task-specific but limited in scalability due to high development costs, generative AI models are widely accessible and adaptable. However, their clinical reliability will depend on rigorous diagnostic accuracy studies, which are essential to validate their sensitivity and specificity in dermoscopic image analysis. Recent research has investigated the application of large language models (LLMs), such as GPT-4 Turbo with Vision, in dermatological diagnostics. Specifically, studies have evaluated their performance in distinguishing between melanomas and benign nevi from dermoscopic images [1, 2]. Initial assessments indicated that GPT-4 Turbo with Vision exhibited lower diagnostic accuracy compared to conventional dermoscopy methods. However, subsequent experiments demonstrated significant improvements by incorporating few-shot learning—a prompt-engineering technique in which task-specific examples are provided directly to the model. For instance, one such experiment reported an initial accuracy of 51.3 ± 1.3% without examples, which improved markedly to 71.1 ± 1.5% when a single example was included. Further enhancements were observed with two or more examples, reaching accuracies of approximately 75–77% [3]. These more advanced prompt engineering methods are intriguing but may be impractical in clinical settings.

Another study has employed custom GPT models, created through OpenAI's GPT editor, to perform narrowly defined tasks, such as identifying dermoscopic criteria of basal cell carcinoma (BCC) and distinguishing BCC from non-BCC images. In this case, while the diagnostic accuracy remained below that of human-driven dermoscopy, the results were more promising. Dermoscopy achieved 96.8% accuracy (98.3% sensitivity, 92.0% specificity) versus the AI model's 84.4% accuracy (94.0% sensitivity, 76.7% specificity) [4].

Trager et al. recently evaluated ChatGPT's ability to diagnose dermoscopy images and suggest management strategies. However, with a limited sample of just 14 lesions—including benign, pre-malignant, and malignant cases—generalizability remains a concern. The model

correctly identified the primary diagnosis in 57% of cases, improving to 86% when considering its top three differentials. Accuracy varied by lesion type: malignant cases were prioritized correctly 63% of the time (100% inclusion in differentials), while benign lesions were accurately identified as the top diagnosis in 40% of cases (60% in differentials). Management recommendations matched expert consensus in 71% of cases, though the model over-recommended biopsies, especially for benign lesions (60%). For malignant cases, management was 100% accurate, but overall biopsy rates (86%) raised concerns about overtreatment. Key limitations include the small, non-diverse sample, lack of standardized protocols, and static testing [5]. Presenting three potential diagnoses also offers limited clinical utility, as it lacks direct applicability in decision-making. This approach is most relevant in primary care settings, where physicians may require additional diagnostic support. Conversely, in higher referral centers, where specialized expertise is readily available, its usefulness is likely minimal unless significant advancements enhance its precision. Furthermore, primary care physicians typically formulate their top differential diagnoses independently and are more likely to seek refined guidance or confirmatory insights rather than a broad diagnostic list.

Karampinis et al.'s study evaluated ChatGPT's ability to interpret and generate differential diagnoses using metaphorical dermoscopic language, confirming its understanding of these clinical terms. The AI excelled in BCC cases but struggled slightly with squamous cell carcinoma (SCC) and inflammatory conditions. Customizing outputs with patient-specific details improved accuracy, highlighting ChatGPT’s flexibility and potential as a diagnostic and educational tool despite some variability in performance [6].

A head-to-head comparison evaluated the performance of Claude 3 Opus and ChatGPT-4 in diagnosing melanoma via dermoscopy. Using 100 histopathology-confirmed images (50 melanomas, 50 benign nevi) from the ISIC archive, both models were given identical prompts to provide their top three differential diagnoses. Accuracy was modest, with Claude achieving 56% accuracy and 54.9% sensitivity, while ChatGPT reached 48% accuracy and 56.86% sensitivity. Claude outperformed ChatGPT in malignant/benign triage. The study concluded that while AI provides some insights, it remains far from replacing trained clinicians—though advancements in this field are progressing at an unprecedented pace [1].

Another study evaluated ChatGPT-4's ability to diagnose dermoscopic images from the PH2 dataset (200 high-resolution images of common nevi, atypical nevi, and melanomas). The model demonstrated moderate melanoma detection accuracy (68.5%) but had limited sensitivity (52.5%) and specificity (72.5%). For identifying suspicious lesions, it achieved 68% accuracy and 78% precision but fell short of expert diagnoses (P = 0.0169). Key limitations included a small dataset from a high-risk melanoma database and a restricted benign differential, limiting generalizability. Researchers highlighted the need for expanded training data and broader validation to improve real-world telehealth applications [7].

Shifai et al. assessed ChatGPT Vision's melanoma detection on 100 confirmed dermoscopic images (50 melanomas, 50 benign nevi). The model showed low accuracy (36%) when considering only its top diagnosis (32% sensitivity, 40% specificity). Performance improved with its top three predictions (56% sensitivity, 53% specificity, 55% accuracy). When distinguishing malignant from benign lesions, accuracy rose to 62%, though sensitivity and specificity varied inversely (top-1: 46%/78%; top-3: 78%/47%). Limitations of this study included small sample size, exclusion of complex cases, and lack of clinical context. However, an interesting finding emerged: its ability to describe dermoscopic features, suggesting potential as a supplementary tool and a foundation for future improvements [2].

Researchers have also explored the ability of LLMs to generate diagnoses based on physician-provided descriptions. While insightful, this approach carries inherent limitations, as the input may be influenced by human bias, potentially affecting the model's diagnostic interpretation [6].

In this study, we adopted a broader approach to more accurately simulate real-world clinical scenarios. Our study focused on evaluating the overall diagnostic accuracy of the general-purpose foundational model GPT-4 Turbo with Vision across the full spectrum of the most common neoplastic dermoscopic findings encountered in routine practice.

In dermatologic diagnostics, a single lesion's differentials can span vastly different pathological categories, with benign and malignant entities often sharing overlapping features. Clinicians navigate this complexity daily, distinguishing between lesions that may appear similar dermoscopically but carry profoundly different prognostic and therapeutic implications.

By including a broad spectrum of diagnoses—benign and malignant, melanocytic and non-melanocytic, keratinocyte-derived and vascular tumors—we ensure AI model faces challenges that mirror real-world practice. A narrow selection of cases would risk oversimplifying the task, inflating AI performance, and limiting its clinical relevance. To more accurately capture the fast pace of primary care, we intentionally limited the inclusion of detailed demographic data and adopted a streamlined "photo-to-answer" approach.

We selected GPT-4 Turbo with Vision due to its market dominance in general-purpose multimodal AI and widespread accessibility via public interfaces. As a continuously evolving model, the protocol also serves as a benchmark for future AI developments. This study assesses its diagnostic accuracy in dermoscopic image interpretation, focusing on key objectives:

Primary Objective

- To assess the overall diagnostic accuracy of GPT-4 Turbo with Vision in dermoscopic image interpretation, the proportion of correct classifications against a reference standard will be evaluated.

Secondary Objectives

- To evaluate diagnostic certainty, including re-prompting frequency and unclassifiable response rates.
- To assess lesion-specific accuracy across different lesion types.
- To evaluate triage performance in classifying lesions as benign, precancerous, or malignant.
- To assess the accuracy of AI-generated management recommendations (excise, follow-up/ treat, no follow-up).
- Assess explainability by having an expert evaluate AI-generated dermoscopic features (up to 4 per image) for accuracy, summarizing results as mean correctness, feature accuracy distribution, and a 95% confidence interval

This study addresses a core question faced by clinicians considering AI-assisted diagnostics: can these systems be reliably applied in routine practice? Whereas previous research has often assessed models under highly controlled conditions using carefully selected datasets limited

to specific diagnostic categories and narrow differentials, real-world implementation requires performance across diverse clinical presentations without such constraints. Our findings seek to narrow this gap by providing evidence on model performance in more representative settings, designed to approximate real-world conditions as closely as possible. The evaluated model was not specifically trained on dermoscopic image datasets, unlike conventional dermatology-specific AI tools, which may limit their baseline diagnostic proficiency in this domain.

In underserved populations, barriers like uninsurance (36.4%), living in medically underserved areas (22.7%), and poverty (33.3%) hinder access to specialized dermatology care [8]. Our study evaluates the potential of ChatGPT-4 Turbo with Vision for dermoscopic image interpretation, which could help guide the broader integration of AI tools to improve referral accessibility in clinical settings.

The study's findings are primarily applicable to primary care settings—such as triage of referrals in general practice or initial assessments in general dermatology clinics—and are intended to inform practice in more specialized environments, such as academic centers where dermoscopy expertise is available. This protocol adheres to the latest STARD 2015 (Standards for Reporting Diagnostic Accuracy Studies) and TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Artificial Intelligence extension) guidelines, ensuring comprehensive and transparent reporting of diagnostic accuracy and AI-based predictive modeling [9, 10].

2 Methods and analysis

2.1 Study design

This study is a retrospective, cross-sectional diagnostic accuracy study evaluating the performance of GPT-4 Turbo with Vision in dermoscopic image interpretation. Using publicly available, pre-annotated dermoscopic image datasets from the International Skin Imaging Collaboration (ISIC) Archive and DermNet New Zealand (DermNetNZ), the study ensures high-quality reference diagnoses without direct patient recruitment or data collection.

A total of 750 images, representing benign and malignant neoplasms, melanocytic and non-melanocytic lesions, and keratinocyte-derived and vascular tumors, will be included.

Each image will be analyzed by GPT-4 Turbo with Vision, and the AI-generated diagnosis will be compared against the reference standard. The model will not have access to clinical history or metadata, simulating real-world usage conditions. No randomization or blinding is applied, as all cases are pre-selected from validated datasets.

2.2 Dataset Composition

The dataset comprises 750 prospectively curated dermoscopic images, selected to mirror real-world diagnostic challenges while ensuring statistical validity. Reflecting prevalence patterns in dermatology practice, we maintained a 5:2 benign-to-malignant/precancerous ratio (533:217 cases). Benign lesions included melanocytic nevi (250 cases), with atypical/ dysplastic variants (75 cases) to stress-test specificity against melanoma mimics, and seborrheic keratosis (SK; 150 cases), including pigmented subtypes to evaluate pigment network discrimination. The classification of dysplastic nevi is contentious. Although dysplastic naevi are technically considered benign for the purposes of dataset composition, owing to their nature and to better reflect prevalence patterns, they were classified as precancerous in the triage analysis due to their clinical urgency and management considerations. Dermatofibroma (30 cases) and pooled "Other Benign" lesions (103 cases)—hemangiomas, angiokeratomas, lichen planus-like keratosis, pyogenic granuloma, sebaceous hyperplasia, neurofibroma, and solar lentigo—ensured diversity without overrepresenting low-prevalence diagnoses. Malignant lesions prioritized melanoma (50 cases), equally stratified into invasive (Breslow 0.5–4.0 mm) and in situ subtypes (including lentigo maligna) to assess AI sensitivity across disease severity. BCC (87 cases) and SCC (36 cases) were included for their clinical urgency and subtype heterogeneity (high-risk [infiltrative, sclerosing/morpheaform] and low-risk [superficial, nodular, fibroepithelial] BCC; invasive [including keratoacanthoma] and in situ [Bowen's disease] SCC). Precancerous actinic keratosis (AK; 44 cases) spanned grades I–III and pigmented variants to evaluate AI's detection of early malignant transformation. AK was categorized separately from malignant lesions in analyses but included in the malignant group for ratio calculations to reflect clinical urgency. Lesion severity was stratified to include early to advanced stages (e.g., in situ vs. invasive melanoma, Grade I–III AK, etc.), ensuring that AI performance is tested across the diagnostic spectrum. Lesion subgroups are aligned with well-established prognostic criteria

[11, 12, 13, 14]. The dataset intentionally incorporates diagnostically challenging mimics to stress-test AI specificity and mirror real-world dilemmas. Diagnostic categories correspond to World Health Organization Classification of Skin Tumours (4th edition) criteria [15]. Consecutive case selection from validated repositories (e.g., ISIC, DermNet) avoids manual preselection bias, ensuring natural difficulty distribution.

Dermoscopy images were sourced from the ISIC Archive (international academic centers in Europe and North America), DermNet NZ (Australasian clinics), and textbooks. The exact number of centers is unavailable, and the dataset includes lesions from both primary and secondary care. The dermoscopic images used in this study will be sourced from the repositories defined above between March 1, 2025 and April 1, 2025. No synthetic data will be analyzed. (Table 1). The demographic details (e.g., age, sex, ethnicity, geographic location) of the individuals represented in the dermoscopic images are not universally available in the defined repositories. As a result, the representativeness of the dataset with respect to specific population subgroups cannot be fully assessed.

GPT-4 Turbo with Vision is a proprietary system, and its development data - including the types, distributions, and sources of dermoscopic images used for training - are not publicly disclosed. This lack of transparency prevents direct comparison between our evaluation dataset and the model's original training data.

Table 1 Datatset composition

Category	Subtype	Cases	% Total	Subgroup Details
Benign (71%)	**Melanocytic nevi**	250	33.3	Common Acquired Nevi: 175 Atypical/dysplastic Nevi: 75
	Seborrheic Keratosis	150	20	N/A
	Dermatofibroma	30	4	N/A
	Other Benign	103	13.7	Hemangiomas: 30 Angiokeratomas: 4 Lichen planus like keratosis: 18 Pyogenic Granuloma: 10 Sebaceous Hyperplasia: 6 Neurofibroma: 7 Solar lentigo: 28
Malignant (23%)	**Melanoma**	50	6.7	Invasive: 25 In Situ: 25

Table 1 Datatset composition

	Basal Cell Carcinoma	87	11.6	High-risk (infiltrative and sclerosing/ morpheaform): 30 Low-risk (superficial, nodular, fibroepithelial): 57
	Squamous Cell Carcinoma	36	4.8	Invasive (incl. keratoacanthoma): 25 In Situ (Bowen's): 11
Precancerous (6%)	**Actinic Keratosis (pigmented and non-pigmented)**	44	5.9	Grade I: 15 Grade II: 15 Grade III: 14

2.3 Eligibility Criteria

In this study, potentially eligible images were identified based on predefined inclusion criteria. The selection process will be conducted by systematically reviewing publicly available and validated dermoscopic image repositories, including the ISIC Archive and DermNet NZ. These sources provide high-quality reference diagnoses, making them scientifically reliable. The dataset will include benign and malignant neoplasms, melanocytic lesions, and non-melanocytic lesions—including keratinocyte-derived and vascular tumors (Table 1)—to reflect real-world diagnostic diversity.

Study inclusion criteria include:

- Dermoscopy images must be sourced from publicly available, validated dermoscopic repositories—such as the ISIC Archive, DermNet NZ, and dermatology textbooks—to ensure consistency with established reference standards.
- Each image must have a verified reference diagnosis, confirmed by these trusted sources, serving as the ground truth for diagnostic accuracy assessment.
- Dermoscopic characteristics (e.g., pigment network, vessels, globules) must be clearly identifiable by the human eye. This will be manually verified by investigator with expertise in dermoscopy to ensure diagnostic utility.
- Dermoscopy images obtained using various photographic equipment will be incorporated to reflect real-world variability and prevent model overfitting to a specific imaging modality.

Study exclusion criteria include:

- Images without a definitive reference diagnosis, including cases with conflicting annotations (e.g., ‘probable BCC’), will be excluded to maintain diagnostic certainty. No imputation will be performed, incomplete cases will be excluded.
- Images that are blurred, contain artifacts, or are degraded by technical flaws (e.g., overexposure, underexposure, glare, shadowing) will be excluded to prevent confounding effects on AI performance, as determined by the investigators.
- Any duplicate images within the dataset will be systematically identified and removed manually to prevent overrepresentation and avoid data redundancy bias.
- Images containing visible text and annotations that could bias AI interpretation will be excluded.

Participant-level eligibility criteria (e.g., age, sex, ethnicity) could not be applied because demographic metadata were unavailable in the repositories. This limits our ability to assess model generalizability across population subgroups. No data on prior treatments (e.g., cryotherapy, biopsies) were available or included, as the AI model interprets images in isolation, reflecting real-world scenarios where treatment history may not be provided with image submissions. To eliminate potential bias from non-visual data, all images were preprocessed prior to AI analysis. Metadata will not be included in the downloads from the databases, and the filenames will already be anonymized at the time of download. This process guaranteed that the AI model analyzed images based solely on pixel content, mirroring real-world scenarios where clinicians submit images without supplemental clinical or technical context.

2.4 Dataset Selection Process

Images were selected consecutively within each category (e.g., melanoma, BCC, nevi) from publicly available, validated repositories in the following order: ISIC Archive (n = 705), DermNet NZ (n = 45), and dermatology textbooks (n = 0), prioritizing larger databases first. Because image selection from the databases was sufficient, we never needed to use the third-tier data source—dermatology textbooks. This approach ensured proportional representation of lesion types and minimized selection bias. Eligible images were included in the order they appeared, with exclusions only for predefined criteria such as lack of a reference diagnosis (n = 0), poor image quality (n = 13), or duplication (n = 35). Duplicate images across

repositories were identified and removed manually to prevent overrepresentation. Final dataset verification was conducted by an investigator to ensure methodological rigor and adherence to study criteria. This consecutive selection strategy supports an objective and unbiased assessment of AI model performance, maintaining real-world applicability. To eliminate potential bias from non-visual data, all images were preprocessed prior to AI analysis. No image harmonization (e.g., resolution thresholds, color calibration, resizing, cropping, contrast enhancement) was performed, reflecting real-world clinical practice where such standardization is typically absent. Reference diagnoses were standardized to subtype-level categories across repositories (e.g., “melanoma” for all melanoma variants, “BCC” for BCC subtypes, “SCC” for SCC subtypes, “AK” “nevi,” “seborrheic keratosis,” “dermatofibroma,” or “other benign” for pooled non-malignant lesions). The AI model (GPT-4 Turbo with Vision) is proprietary with undisclosed training data, making it impossible to confirm prior exposure to our dataset. This is a potential limitation, but as general-purpose model without dermatology-specific fine-tuning, their performance reflects real-world use.

2.5 Index test

We evaluated the latest publicly available version of GPT-4 Turbo with Vision (OpenAI) at the time of data collection, using default configurations to generate single-attempt outputs per image. This approach mirrors how clinicians typically interact with such tools in real-world settings. The model produced categorical diagnostic classifications (e.g., "melanoma," "benign nevus") rather than probabilistic outputs.

This benchmark study assesses the AI model version available during our data collection period (June 2–6, 2025), establishing a performance baseline. Given the architectural stability of multimodal systems, significant declines in accuracy are unlikely. We anticipate future performance improvements through continued training on medical data and version inheritance. Our protocol includes standardized metrics to enable longitudinal tracking of diagnostic stability.

2.6 Prompting Strategy and Standardization

The AI model was evaluated using a standardized, structured prompt designed to elicit optimal diagnostic reasoning while minimizing suggestive bias. The prompt was crafted to

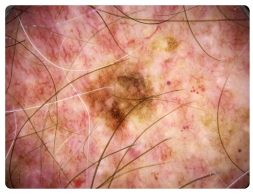
simulate clinical case presentation, emphasizing objective analysis of dermoscopic features without leading toward specific diagnoses. For each case, the model received only a single dermoscopic image without clinical history or contextual data. No definitions of dermoscopic features were provided. The chat history was cleared before each submission to prevent carryover effects from previous cases. The standardized prompt used for each model was as follows:

"You are a dermatology expert trained in dermoscopic image analysis. Below is a dermoscopic image of a skin lesion. Analyze its features, including color, structure, symmetry, and characteristic dermoscopic patterns. List the top features influencing your diagnosis, then provide:

- The single most likely diagnosis.
- Triage category: benign, precancerous, or malignant.
- Clinical management: Excise, Follow-up/Treat, or No follow-up.“

This prompt ensures that AI-generated responses are interpretable and comparable, enabling a structured assessment of diagnostic accuracy. An example of the dialog window is presented in figure 1. Our approach was designed to balance cutting-edge prompt design effectiveness with real-world clinical applicability. While more advanced strategies may be less practical in daily medical settings, certain approaches remain both effective and easy to implement. To optimize diagnostic accuracy and usability, we employed Role Prompting, Instruction Prompting, Output Formatting, Contextual Prompting, and Specificity and Clarity—ensuring

ChatGPT 4o



You are a dermatology expert trained in dermoscopic image analysis. Below is a dermoscopic image of a skin lesion. Analyze its features, including color, structure, symmetry, and characteristic dermoscopic patterns. List the top features influencing your diagnosis, then provide:
The single most likely diagnosis.
Triage category: benign, precancerous, or malignant.
Clinical management: Excise, Follow-up/Treat, or No follow-up.

Dermoscopic Image Analysis:

1. Color:

- Multiple colors are present, including:
  - Light brown
  - Dark brown
  - Reddish-pink (vascular background)
  - Faint yellowish hues

2. Structure:

- **Asymmetry**: The lesion appears asymmetric in both color and structure.
- **Pigment network**: Irregular and atypical, not clearly net-like.
- **Blotches**: Presence of multiple irregular blotches and globules, especially in the center.
- **Regression structures**: Whitish areas and peppering (gray dots) suggest regression.
- **Vascular pattern**: Polymorphous vessels, including dotted and linear irregular vessels.
- **Follicular and perifollicular structures**: Absent or non-specific.

3. Symmetry:

- The lesion shows **asymmetry in both structure and color**, which is concerning.

4. Characteristic Dermoscopic Patterns:

- Irregular blotches and pigmentation
- Regression structures (peppering and scar-like depigmentation)
- Polymorphous vessels
- Lack of a uniform pigment network

Top Features Influencing Diagnosis:

1. **Asymmetry** in structure and color
2. **Multiple colors**, especially dark blotches with regression areas
3. **Regression structures** (white areas and peppering)
4. **Polymorphous vessels**
5. **Irregular blotches and structureless zones**

Diagnosis & Recommendations:

- **Most Likely Diagnosis**: **Superficial spreading melanoma**
- **Triage Category**: **Malignant**
- **Clinical Management**: **Excise**

Rationale: The combination of asymmetry, regression structures, atypical vascular patterns, and multiple colors strongly favors melanoma over benign entities like nevi or seborrheic keratoses.

If desired, I can assist with generating a report or recommending further imaging/follow-up.

2/2

Figure 1. *A sample dialog window illustrating the user interface.*

expert-level analysis, structured feature assessment, and clear, reproducible conclusions. More complex methods, such as Few-Shot Prompting and Chain of Thought Reasoning, were excluded because they rely on extensive examples or intricate reasoning, which may not be feasible in routine clinical practice [16].

In this study, indeterminate index test results were handled by re-prompting the AI model to ensure a definitive response (Figure 2). A standardized follow-up prompt was used:

"Give me one single best diagnosis and classify the lesion as benign, precancerous, or malignant."

This approach is designed to align with how physicians would intuitively interact with AI tools in practice, seeking definitive diagnoses to inform decision-making, thereby ensuring AI outputs are clear and actionable. Unclassifiable responses are defined as instances where the AI model fails to provide a definitive, clinically actionable diagnosis after standardized re-prompting. This includes: [A] ambiguous outputs (e.g., “possibly benign or malignant”) and [B] refusal to diagnose (e.g., “insufficient information”). In the primary analysis, unclassifiable responses are treated as incorrect (malignant/precancerous cases → false negatives; benign cases → false positives) to reflect real-world clinical utility, where indeterminate answers cannot guide management. Sensitivity analyses excluded unclassifiable cases to assess their impact on accuracy metrics. Unclassifiable rates were tracked per lesion type to identify patterns (e.g., clustered uncertainty in early melanomas or rare tumors), offering insights into AI limitations. Fig. 1 provides a detailed depiction of the flow chart of the study, illustrating the sequential steps, processes, and methodologies undertaken throughout the research. For the other outcomes, re-prompting will not be conducted, and any insufficient or missing responses will be automatically considered incorrect.

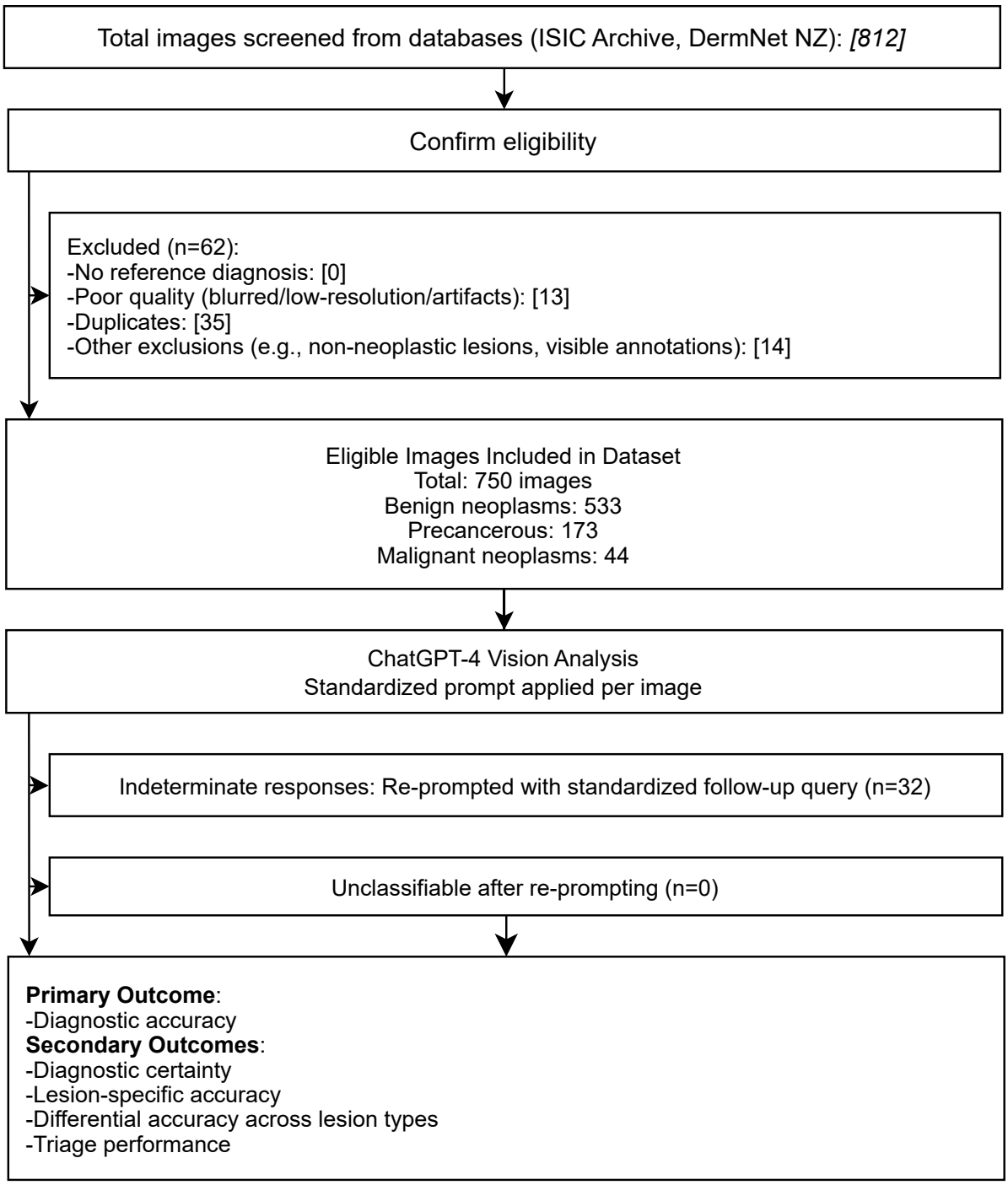


Figure 2 *Flow Chart of Study: Diagnostic Accuracy of GPT-4 Turbo with Vision in Dermoscopic Image Analysis*
ISIC, International Skin Imaging Collaboration
The study begins with image identification and screening, sourcing [812] images from validated databases, including the ISIC Archive, DermNet NZ, and dermatology textbooks. Exclusions include images without a reference diagnosis, poor-quality images, duplicates, visible annotations, and non-neoplastic lesions. The final dataset consists of 750 eligible images, distributed across benign and malignant neoplasms, as well as melanocytic and non-melanocytic lesions.
GPT-4 Turbo with Vision analyzes each image using a standardized prompt. Indeterminate responses are re-prompted, and unclassifiable images are recorded. Diagnoses are compared against pre-annotated reference standards from validated sources. This study adheres to STARD 2015 guidelines, ensuring methodological transparency and diagnostic outcome reporting.

2.7 Defining "Correct" and "Incorrect" Diagnosis

Diagnostic accuracy was assessed at the lesion-type level, with an AI-generated diagnosis considered "correct" if it matches the reference standard's primary pathological category ("BCC" for any BCC subtype, "melanoma" for all melanoma variants, "SCC" for any SCC subtype, "actinic keratosis," "nevi," "seborrheic keratosis [SK]," "dermatofibroma," or "Other Benign" for all defined benign lesions). Responses were classified as "incorrect" if they mismatch the lesion type (e.g., misdiagnosing melanoma as a nevus) or remain unclassifiable after standardized re-prompting.

AI-generated diagnoses that include voluntary subtype specifications (e.g., "nodular BCC") were be coded exclusively at the lesion-type level (e.g., "BCC"), with subtype labels disregarded in all analyses. This approach ensures consistency with the study's focus on clinically relevant distinctions, such as differentiating benign from malignant lesions, while minimizing bias introduced by statistically underpowered or uninterpretable subtype-level errors. Pilot testing demonstrated a low rate of subtype-specific diagnoses (10% of responses), making such analysis uninformative within this protocol. Moreover, assessing the accuracy of voluntarily provided subtype diagnoses would introduce self-selection bias, as it would likely overestimate AI performance. Our exploratory pilot analysis suggests that AI predominantly assigns subtype labels when the lesion exhibits highly characteristic features of that subtype, thereby excluding clinically equivocal cases.

Although lesion identification outcomes were grouped into broad "lesion-type" categories, one subtype required a tailored approach. For dysplastic nevi, a diagnosis was considered correct if it identified any melanocytic nevus (benign or dysplastic). However, for triage performance and management evaluation, responses were only considered correct if dysplastic nevi were classified as precancerous and the recommended management was excision, to reflect their clinical urgency more accurately.

The "Other Benign" category (n=103) comprises pooled low-prevalence benign lesions (e.g., hemangiomas, angiokeratomas, trichoepitheliomas) to ensure dataset diversity without overrepresenting rare entities. For lesions in the "Other Benign" category (n=103), a classification was considered correct if the AI identifies the lesion as benign, even if the specific subtype is misclassified (e.g., pyogenic granuloma vs. trichoepithelioma). This aggregated approach reflects clinical priorities, as these lesions are generally managed

similarly (excision or observation), with ruling out malignancy taking precedence over pinpointing the exact benign subtype. It also ensures statistical stability, given the small sample sizes of individual subtypes. While this method aligns with real-world diagnostic workflows, subtype misclassifications could obscure rare cases where precise benign diagnosis influences management (e.g., lesions with malignant potential). Due to clinical heterogeneity and small sample sizes within subgroups, this category was not analyzed by individual lesion type. Instead, it was included in aggregate benign lesion calculations (e.g., specificity) to reflect real-world diagnostic scenarios where AI must broadly distinguish benign from malignant lesions. A management recommendation was considered correct only if the diagnosis matched the reference standard and the recommended management category aligned with the reference standard's urgency (malignant → excise; precancerous → follow-up/treat; benign → no follow-up). Otherwise, it was considered incorrect, recognizing that, in real-life practice, therapeutic decisions often involve nuances beyond the excise/follow-up/treat/no follow-up framework, depending on the specific lesion diagnosis.

For the explainability analysis, a single dermoscopy expert will review AI-generated feature lists and images, determining for each case how many of the cited features (0 to 4) exists in the image and align with the reference diagnosis.

2.8 Outcomes and statistical design

The primary outcome is the diagnostic accuracy of GPT-4 Turbo with Vision in dermoscopic image interpretation. This will be measured by calculating overall accuracy. Model's single best diagnosis per case will be collapsed into binary form and classified as "correct" or "incorrect" against the reference standard. The classified groups will include melanoma, BCC, AK, nevi, seborrheic keratosis, SCC, dermatofibroma, and the pooled "Other Benign"category (see Defining "Correct" and "Incorrect" Diagnosis). Overall accuracy will be reported as the proportion of correct classifications, with 95% confidence intervals (CIs) estimated via bootstrapping (10,000 resamples).

Secondary outcomes include:

2. Diagnostic certainty: Re-prompting frequency (i.e., the number of attempts to elicit a classifiable response) and rates of unclassifiable responses will be reported globally.

3. Lesion-Specific Diagnostic Performance:

For each lesion type (melanoma, BCC, SCC, AK, nevi, seborrheic keratosis, dermatofibroma):

- Sensitivity measures the AI's ability to correctly identify true cases (e.g., melanoma sensitivity = the proportion of melanomas accurately diagnosed out of all melanomas).
- Specificity assesses the AI's ability to exclude non-target lesions (e.g., melanoma specificity = the proportion of non-melanomas correctly classified as non-melanoma).

  To contextualize performance, specificity is analyzed in two ways:

  A. Broad specificity: Includes all non-target lesions (e.g., melanoma vs. BCC, AK, and benign lesions). This may inflate metrics due to visually distinct cases (e.g., melanoma vs. hemangioma).

  B. Mimic-specific specificity: Focuses on lesions with overlapping features, reflecting real-world diagnostic challenges: Melanoma vs. dysplastic nevi (n=50 vs. 75), critical for avoiding missed malignancies. SCC vs. AK (n=36 vs. 44), guiding management escalation. BCC vs. benign adnexal tumors (n=87 vs. 13), reducing over-excision.

  This study employed a lenient rule for specificity analysis, wherein any ChatGPT diagnosis other than the target lesion (e.g., melanoma) was considered a correct exclusion, regardless of diagnostic accuracy for alternative conditions. While this approach reflects real-world clinical scenarios—where an incorrect but non-malignant diagnosis may still avert unnecessary intervention—it may overestimate specificity by classifying misclassified benign or unrelated lesions (e.g., a nevus labeled as 'BCC') as true negatives. Cohen's Kappa (κ) quantifies diagnostic agreement beyond chance. Mimic-specific analyses prioritize high-stakes lesion pairs (e.g., melanoma vs. dysplastic nevi, SCC vs. AK, BCC vs. benign adnexal tumors) to balance statistical feasibility and clinical urgency. Smaller subgroups ($n < 20$) are flagged as exploratory. The pooled "Other Benign" category contributes to broad specificity but is excluded from individual analyses. Confidence intervals for binomial proportions were calculated using Wilson score intervals.

4. Triage performance (benign, precancerous, malignant): This will be evaluated using a 3×3 contingency table. Sensitivity and specificity per category, as well as overall accuracy (total correct classifications), will be calculated. Weighted Cohen's Kappa (κ) will quantify agreement, assigning higher penalties to clinically significant errors: benign/malignant

discordance (weight = 1.0) will be penalized more than benign/precancerous discordance (weight = 0.5). Weighted κ prioritizes clinically critical misclassifications, unlike simple accuracy, which treats all errors equally. This approach ensures the analysis reflects real-world diagnostic risks while maintaining statistical rigor.

We acknowledge that results for SCC (n = 36) and dermatofibroma (n = 30) should be interpreted cautiously due to modest sample sizes. If so, we will include Wilson score CIs to mitigate small-sample bias.

5. Management accuracy: Overall percentage of cases where both diagnosis and management recommendation are correct (95% CI).

6. For the exploratory explainability analysis, a single dermoscopy expert will review each AI-generated feature list (up to 4 features per case) and corresponding image, assessing whether each cited feature is present and aligns with the reference diagnosis. The number of correct features per case (0–4) will be aggregated to report: (1) the mean proportion of correct features across cases, (2) the distribution of correct features (e.g., % of cases with 0, 1, 2, 3, or 4 correct features), and (3) a 95% confidence interval for the overall accuracy.

2.9 Sample Size

The sample size of 750 dermoscopic images was determined through precision-based calculations to ensure clinically informative confidence intervals (CIs) for diagnostic accuracy. For the primary outcome (overall accuracy), we assumed maximal variance ($p = 0.5$) to calculate a worst-case margin of error:

$$n = \frac{Z_{\alpha/2}^2 p(1-p)}{e^2} = \frac{1.96^2 \cdot 0.25}{0.05^2} = 385.$$

By increasing the sample size to 750 images, we obtain a 95% confidence interval with a ±5% margin of error around an expected accuracy of 80%. This design yields approximately 95% power to detect a 5% difference in accuracy [17]. It should be acknowledged that the analyses of SCC and dermatofibroma are underpowered and should be considered exploratory.

The study evaluates diagnostic accuracy at the lesion-type level (e.g., melanoma, BCC, SCC, AK, nevi, seborrheic keratosis, dermatofibroma, and pooled “Other Benign” lesions). Importantly, the standardized diagnostic prompt was designed to elicit diagnoses at this lesion-type level, not to distinguish subtypes (e.g., invasive vs. in situ melanoma,

pigmented vs. non-pigmented AK, or AK grades I–III). While subtype-level annotations (e.g., melanoma invasiveness, AK severity) exist in the dataset, these will not be analyzed due to insufficient statistical power and lack of alignment with the prompt's diagnostic scope. Any observed trends in subtype performance (e.g., differences between invasive and in situ melanoma) are exploratory, as the study was not powered to detect statistically significant differences at this granularity. As this study evaluates pre-existing AI model, the entire sample (n=750) was used for performance evaluation. The sample size of 750 images ensures a precision of ±3.5% for the primary outcome, but it was not optimized for all subgroup analyses.

3 Results

The study workflow is detailed in figure 2. From an initial pool of 812 screened images, we excluded 62 (7.6%) for quality concerns: 13 (1.6%) showed poor image quality (e.g., blurring or artifacts), 35 (4.3%) were duplicates, and 14 (1.7%) had incorrect labeling. The final analytical dataset comprised 750 images, the exact composition of the dataset is detailed in table 1. No exclusions were made for missing reference diagnoses. The study employed OpenAI's GPT-4 Turbo with Vision (May 2024 version) accessed exclusively via cloud application programming interface.

This study did not account for or analyze data clustering (e.g., by demographic factors, geographic location, or clinical center). All lesions were evaluated as independent cases without adjustment for potential clustering effects. Consequently, we do not report heterogeneity in model performance across clusters.

3.1 Diagnostic Accuracy and Certainty of GPT-4 Turbo with Vision

GPT-4 Turbo with Vision demonstrated an overall diagnostic accuracy of 37.6% (282 of 750; 95% CI 34.1–41.1) in classifying dermoscopic images, as estimated by non-parametric bootstrap resampling (10 000 iterations). Diagnostic accuracy was defined as the proportion of correct diagnoses relative to histopathological or expert confirmation, with exact concordance assessed for eight primary lesion categories: BCC, melanoma, SCC, AK, nevi, seborrhoeic keratosis, dermatofibroma, and other benign lesions. Subtype distinctions were excluded, and all diagnoses were mapped to the lesion-type level. For images classified as

other benign lesions (n=103), any predefined ‘other benign’ designation was accepted as correct; one discordant case occurred (1 of 750, 0.13%). Follow-up queries were required for 32 cases (4.3%, 95% CI 3.0–5.9), with no unclassifiable images (0.0%, 95% CI 0.0–0.5). The mean recall score across all diagnostic classes was 0.34 (95% CI 0.30–0.37; bootstrap, 10 000 resamples), calculated with equal weighting for each lesion type.

3.2 Lesion-Specific Sensitivity & Specificity for each lesion type

As one of the predefined secondary endpoints, the analysis evaluated lesion-specific sensitivity and specificity for each diagnostic category. Sensitivity was defined as the proportion of correctly identified target lesions, whereas specificity reflected the proportion of correctly excluded non-target lesions. Specificity was assessed using two complementary approaches: broad specificity, in which all non-target lesions were considered true negatives, and mimic-specific specificity, which restricted true negatives to clinically and visually similar lesions. The mimic sets included melanoma versus dysplastic melanocytic nevi, SCC versus AK, and BCC versus the ‘other benign’ category, which comprised mimics such as sebaceous hyperplasia, lichen planus-like keratosis, angiokeratomas, and hemangiomas.

The model exhibited substantial variation in sensitivity across lesion types (Table 2), with the lowest performance observed for AK (0.07, 95% CI: 0.02–0.18) and the highest for melanoma (0.64, 95% CI: 0.50–0.76). Broad specificity—defined as the correct exclusion of all non-target lesions—was ranging from 0.85 (95% CI: 0.81–0.88) for melanocytic nevi to 0.98 (95% CI: 0.96–0.99) for SCC and AK (Table 2). Agreement between model predictions and reference diagnoses, measured by Cohen’s kappa, was highest for melanocytic nevi (κ = 0.45, 95% CI: 0.38–0.52) and BCC (κ = 0.36, 95% CI: 0.26–0.46), while SCC (κ = 0.17, 95% CI: 0.08–0.32) and AK (κ = 0.07, 95% CI: 0.03–0.18) showed minimal concordance.

To assess performance in clinically challenging scenarios, we evaluated mimic-specific specificity for three diagnostically ambiguous pairs (Table 3-5). In the melanoma versus atypical melanocytic nevi comparison (Table 3), specificity was 0.80 (95% CI: 0.70–0.87) for melanoma and 0.80 (95% CI: 0.67–0.89) for nevi, with moderate agreement (κ = 0.44, 95% CI: 0.28–0.61 for melanoma; κ = 0.50, 95% CI: 0.35–0.65 for nevi). For SCC versus AK (Table 4), specificity reached 0.86 (95% CI: 0.73–0.94) for SCC and 0.97 (95% CI: 0.86–1.00) for AK, though agreement was negligible (κ = 0.03, 95% CI: −0.14 to 0.20 for

SCC; κ = 0.04, 95% CI: −0.05 to 0.12 for AK). In the BCC versus benign lesions comparison (Table 5), specificity was 0.90 (95% CI: 0.83–0.95) for BCC and 1.00 (95% CI: 0.96–1.00) for benign lesions, with fair agreement for BCC (κ = 0.34, 95% CI: 0.22–0.46) but limited for benign lesions (κ = 0.20, 95% CI: 0.12–0.28).

Table 2. Lesion-Specific Sensitivity, Broad Specificity, and Agreement with Reference Standard

Lesion type	Correct diagnosis decision		Sensitivity (95% CI)	Cohen's kappa	Broad specificity (95% CI)
Melanocytic Nevi	Not case	424/500	0.59 (0.53–0.65)	0.448	0.85 (0.81–0.88)
	Case	147/250			
Seborrheic keratosis	Not case	560/600	0.18 (0.13–0.25)	0.143	0.93 (0.91–0.95)
	Case	27/150			
Dermatofibroma	Not case	692/720	0.47 (0.30–0.64)	0.359	0.96 (0.94–0.97)
	Case	14/30			
Other Benign	Not case	616/647	0.16 (0.10–0.24)	0.139	0.95 (0.93–0.97)
	Case	16/103			
Melanoma	Not case	489/700	0.64 (0.50–0.76)	0.121	0.70 (0.66–0.73)
	Case	32/50			
Basal cell carcinoma	Not case	617/663	0.43 (0.33–0.53)	0.363	0.93 (0.91–0.95)
	Case	37/87			
Squamous cell carcinoma	Not case	697/714	0.17 (0.08–0.32)	0.172	0.98 (0.96–0.99)
	Case	6/36			
Actinic keratosis	Not case	694/706	0.07 (0.02–0.18)	0.074	0.98 (0.97–0.99)
	Case	3/44			

Table 3. Mimic-Specific Specificity and Cohen's Kappa for Melanoma vs. Atypical Melanocytic Nevi

Lesion type	Total non-target	True negative	False positive	Specificity (95% CI)	Kappa (95% CI)
Melanoma	75	60	15	0.80 (0.70–0.87)	0.44 (0.28–0.61)
Melanocytic Nevi	50	40	10	0.80 (0.67–0.89)	0.50 (0.35–0.65)

Table 4. Mimic-Specific Specificity and Cohen's Kappa for Squamous Cell Carcinoma vs. Actinic Keratosis

Lesion type	Total non-target	True negative	False positive	Specificity (95% CI)	Kappa (95% CI)

SCC	44	38	6	0.86 (0.73–0.94)	0.03 (-0.14–0.20)
Actinic keratosis	36	35	1	0.97 (0.86–1.00)	0.04 (-0.05–0.12)

Table 5. Mimic-Specific Specificity and Cohen's Kappa for Basal Cell Carcinoma vs. Other benign

Lesion type	Total non-target	True negative	False positive	Specificity (95% CI)	Kappa (95% CI)
BCC	103	93	10	0.90 (0.83–0.95)	0.34 (0.22–0.46)
Other benign	87	87	0	1.00 (0.96–1.00)	0.20 (0.12–0.28)

3.3 Management Recommendation Accuracy

Management recommendation accuracy was defined as concordance with the reference standard for both diagnostic classification and management decision (excision, follow-up, or no intervention). For dysplastic nevi, diagnostic concordance was assigned when classified as any melanocytic nevus; management concordance required assignment as precancerous with excision recommended. The model provided concordant recommendations in 208 of 750 cases, yielding an accuracy of 27.7% (95% CI 24.8–30.9; bootstrap resampling, 10 000 iterations).

3.4 Triage Classification Performance

The dataset for triage performance calculation comprised 750 dermoscopic dermatologic lesions with histopathological or expert consensus confirmation, including 400 benign (53.3%), 119 precancerous (15.9%), and 231 malignant cases (30.8%). The artificial intelligence model achieved an overall classification accuracy of 48.8% (366/750; 95% CI 45.2–52.4), with a weighted Cohen's kappa of –0.20 (95% CI –0.26 to –0.14), accounting for the clinical severity of misclassifications through differential weighting (benign–malignant discordance weight = 1.0; benign–precancerous weight = 0.5).

For malignant lesions, the model correctly identified 178 of 231 cases (sensitivity 0.77; 95% CI 0.70–0.83), while misclassifying 40 cases as either benign (n = 32) or precancerous (n = 8). Specificity for malignant lesions was 0.63 (95% CI 0.59–0.67). Precancerous lesions showed the lowest detection rate, with only 13 of 119 cases correctly identified (sensitivity 0.11; 95% CI 0.07–0.18). Specificity for precancerous lesions was 0.89 (95% CI 0.87–0.92). Among benign lesions, 220 of 400 cases were correctly classified

(sensitivity 0.48; 95% CI 0.43–0.53), while 179 were misclassified as malignant and 59 as precancerous. Specificity for benign lesions was 0.65 (95% CI 0.59–0.70). The positive predictive value for benign classification was 0.68 (95% CI 0.62–0.73), and the negative predictive value was 0.45 (95% CI 0.40–0.50). The confusion matrix for AI-based triage classification and the sensitivity and specificity by triage category are presented in Table 6 and Table 7, respectively.

Table 6. Confusion Matrix for AI-Based Triage Classification (Benign, Precancerous, Malignant)

	Predicted		
Reference	Benign	Precancerous	Malignant
Benign	220	59	179
Precancerous	71	13	35
Malignant	32	8	133

Table 7. Sensitivity and Specificity by Triage Category

Category	Sensitivity (95% CI)	Specificity (95% CI)
Benign	0.48 (0.43–0.53)	0.65 (0.59–0.70)
Precancerous	0.11 (0.06–0.18)	0.89 (0.87–0.92)
Malignant	0.77 (0.70–0.83)	0.63 (0.59–0.67)

3.5 Explainability Analysis (Feature-Level Accuracy)

The AI model's explainability was evaluated by quantifying the number of dermoscopic features correctly identified per lesion, as verified by expert dermatologists. Across all cases, the model correctly identified a mean of 3·17 (SD 0.89; 95% CI 3.08–3.25) features per lesion. Feature identification accuracy analysis revealed 454 cases (60.5%) with all four features correctly identified, compared with 48 cases (6.4%) for both zero and one correct features. Intermediate accuracy levels showed 112 cases (14.9%) with three correct features and 88 cases (11.7%) with two correct features.

Performance varied by lesion type (Table 8). The highest mean feature accuracy was observed for melanocytic nevi (mean 3.82, SD 0.75) and melanoma (mean 3.76, SD 0.52). In contrast, the lowest accuracy was recorded for AK (mean 2·02, SD 1·68) and SCC (mean 2.28, SD 1.23), both of which also showed the greatest variability. Intermediate performance

was seen in BCC (mean 3.26, SD 1.03), dermatofibroma (mean 3.00, SD 1.14), and seborrheic keratosis (mean 2.89, SD 1.28), while other benign lesions demonstrated moderate accuracy (mean 2.48, SD 1.30).

Table 8. Mean number of features identified

Lesion Type	Mean (SD)
Actinic keratosis	2.02 (1.68)
Basal cell carcinoma	3.26 (1.03)
Dermatofibroma	3.00 (1.14)
Melanocytic nevi	3.82 (0.75)
Melanoma	3.76 (0.52)
Other benign lesions	2.48 (1.30)
Squamous cell carcinoma	2.28 (1.23)
Seborrheic keratosis	2.89 (1.28)

4 Discussion

Recent advances in AI highlight the potential of foundational LLMs to achieve postgraduate-level expertise across diverse domains, including medicine [18, 19]. However, while their natural language processing capabilities are well-documented, their multimodal proficiency—particularly in visually oriented tasks like dermatologic diagnosis—remains inadequately explored. This retrospective diagnostic study evaluated GPT-4 Turbo Vision (gpt-4-turbo-2024-04-09) using a dataset of 750 dermoscopic images, selected consecutively from ISIC archive and DermnetNZ.

In this study, we address a core question faced by clinicians considering AI-assisted diagnostics: Can generic foundational LLMs be reliably applied to routine dermoscopic diagnosis? To explore this, we designed the study to evaluate model performance in settings that closely reflect real-world practice, incorporating broad datasets, multiple differential diagnoses, representative disease prevalence, and a simplified "photo-to-answer" workflow without complex prompting. This study provides the first 8-class evaluation of GPT-4 Turbo Vision on such a broad spectrum of cutaneous neoplastic lesions, using rigorously annotated

dermoscopic images. The question is particularly urgent because, despite the current lack of robust clinical validation, clinicians are increasingly adopting these systems, often influenced by the enthusiasm and perceived success of LLMs in other domains. However, such premature reliance on LLMs' diagnostic capabilities—driven more by hype than by evidence—risks fostering false trust, setting a potentially hazardous trajectory in the absence of rigorous empirical scrutiny.

Our evaluation demonstrates that GPT-4 Turbo Vision's diagnostic performance in dermoscopic image interpretation—while showing measurable pattern recognition (37.6% overall accuracy vs 16.7% prevalence-weighted chance)—remains clinically inadequate. The model's performance lags significantly behind that of expert clinicians, who achieve 92% sensitivity in melanoma detection [20], a critical deficit given the potentially fatal consequences of missed malignancies.

The balanced multiclass accuracy of GPT-4 Turbo with Vision in this study (mean recall score: 0.34; 95% CI: 0.30–0.37) demonstrates significant limitations compared to dedicated dermatological AI systems. These findings contrast sharply with results from the recent multicenter clinical trial by Menzies et al. (2024), which evaluated specialized mobile-powered AI for pigmented lesion diagnosis in specialist care settings. In their prospective study, a specialized 7-class AI algorithm achieved markedly superior diagnostic accuracy (mean recall: 65.9%; 95% CI: 26.9–100), performing comparably to dermatology specialists (73.8%) while significantly outperforming novice clinicians (35.5%) [21]. This performance gap is likely due to a fundamental difference in system design: the 7-class AI was extensively trained on dermoscopic images, while GPT-4's general-purpose architecture lacks similar optimization for dermatology-specific tasks.

Nonetheless, it's worth noting that the field continues to advance rapidly, with each iteration of ChatGPT demonstrating meaningful performance gains. To illustrate this progression, we compared our findings with those reported by Shifai et al., who evaluated an earlier version of GPT-4V (September 25, 2023) on a simplified dataset comprising 50 melanomas and 50 unequivocally benign (non-atypical) nevi—an artificial diagnostic scenario likely to inflate specificity [2]. In contrast, our study assessed the newer GPT-4 Turbo with Vision on a more clinically realistic dataset: 50 melanomas among 700 diverse benign lesions, including a focused evaluation against 75 dysplastic nevi—the most

challenging melanoma mimics. While the earlier model achieved only 32% sensitivity and 40% specificity in the simplified setting, our more rigorous evaluation revealed substantial gains, with GPT-4 Turbo reaching 64% sensitivity and 80% mimic-specific specificity for melanoma detection. Despite these advancements, the moderate agreement in the melanoma category ($\kappa = 0.44$) underscores that GPT-4 Turbo remains insufficient for diagnostic use.

Interestingly, the model's mimic-specific specificity for melanoma (0.80) surpasses its overall specificity (0.70), suggesting improved discrimination precisely in the most diagnostically challenging scenarios. We hypothesise that for broad classification tasks, GPT-4 Turbo may rely on heuristic "shortcut learning" or the "Clever Hans effect," overemphasising superficial cues such as colour. In contrast, when faced with high-stakes differentials—such as melanoma versus dysplastic nevi—it appears to engage more robust pattern-recognition capacity. This interpretation aligns with prior evidence demonstrating shortcut learning in clinical AI, where models often exploit spurious correlations instead of deeper, context-sensitive features [22]. Additionally, cost-effectiveness considerations may influence such behaviour: a general-purpose model like GPT-4 is designed to optimise utility across tasks and might selectively allocate computational resources or attention to high-risk cases, such as potential melanoma diagnoses. Nevertheless, its performance still falls short of clinical-grade standards (≥ 0.95) [23].

In the NMSC and AK categories, performance was notably poorer. Sensitivity was low across the board: 0.43 for BCC, 0.07 for AK, and 0.17 for SCC. The observed high overall specificity—ranging from 0.93 for BCC to 0.98 for both SCC and AK—likely reflects an artefact of the multi-class classification setting, where false positives are distributed across unrelated categories, thus inflating specificity values. When examining mimic-specific specificity between SCC and AK, specificity dropped to 0.86—still high—despite minimal agreement beyond chance (Cohen's $\kappa = 0.03$ for SCC and 0.04 for AK). This discrepancy suggests that the model's outputs in these cases are shaped more by consistent misclassification patterns than by genuine diagnostic understanding. Rather than addressing real-world clinical challenges—such as differentiating SCC from AK—the model often produced implausible predictions, such as labelling non-pigmented AK lesions as dermatofibroma or melanoma. These errant classifications reveal that high aggregate specificity may conceal critical diagnostic failures. Therefore, relying solely on standard

performance metrics risks overstating the model's clinical utility. Future research should emphasize fine-grained error analysis to elucidate the model's underlying decision processes more accurately.

The model's triage performance—arguably its most clinically relevant function—reveals critical limitations that would be unacceptable in practice. Our results show the AI achieved only 49% overall triage accuracy, with particularly concerning performance gaps: while it correctly identified 77% of malignant lesions (still missing 23% of cancers), it detected a mere 11% of precancerous lesions - a dangerous shortcoming for early cancer prevention. The model's weighted kappa ($\kappa = -0.20$) revealed systematically harmful misclassifications, frequently downgrading high-risk lesions to benign categories. These weaknesses become even more apparent when compared to human capabilities: a study of dermoscopy-naive primary care providers showed their diagnostic accuracy improved from 76.4% to 90.8% after brief training [24]. These findings highlight a key distinction: while primary care physicians can rapidly acquire and retain effective triage skills through focused training, current general-purpose LLMs lack the clinical reasoning required for safe and reliable use in dermatologic triage. Instead of placing premature trust in unvalidated LLMs, efforts should prioritise accessible, evidence-based education—ensuring that clinicians are equipped to make informed decisions without becoming reliant on tools not yet fit for clinical deployment.

The explainability analysis revealed a paradoxical finding: while GPT-4 Turbo with Vision demonstrated poor diagnostic accuracy (37.6%), its ability to identify dermoscopic features was notably better (mean of 3.17 correct features per lesion). This discrepancy suggests that the model's failure lies not in detecting morphological patterns but in integrating them into precise diagnostic conclusions. The particularly poor performance in NMSCs, such as AK (mean 2.02 features) and SCC (mean 2.28), may stem from several factors: (1) NMSCs often exhibit subtler or overlapping dermoscopic features compared to melanocytic lesions; (2) GPT-4's training likely prioritized melanoma over NMSCs due to its higher morbidity; and (3) the model's generalist architecture lacks domain-specific fine-tuning for keratinocyte-derived lesions.

Interestingly, the AI predominantly used non-metaphorical, broad descriptors (e.g., "lines," "dots," and "areas") rather than metaphorical language often employed by human

dermatologists (e.g., "spoke-wheel areas" or "strawberry pattern"). While standardized terminology aids transparency and human pattern recognition, LLMs—with their near-unlimited computational memory and processing capacity—may actually benefit from more granular, literal descriptions that avoid the ambiguity of figurative language. Moving forward, refining the balance between granular descriptors and diagnostic synthesis will likely be important to unlocking AI's potential in dermatology.

Our explainability analysis is exploratory and underscores the need for standardized validation frameworks for explainability in LLMs, particularly to mitigate the "black box" effect. To our knowledge, this is the first study to analyze explainability in generic LLM-based dermoscopy analysis. We did not predefine criteria for feature identification [25], which may have resulted in an overestimation of correctness by accepting overly broad descriptors —highlighting a potential methodological limitation. Future work should aim to establish consensus on whether such fundamental terms are sufficient, or whether additional specificity (e.g., spatial arrangement, colour modifiers) is necessary for clinical relevance.

Our study has key limitations. The dataset underrepresented darker skin tones (Fitzpatrick IV–VI) and rare malignancies, potentially exacerbating real-world performance disparities. While the sample size (n=750) was substantial, low-prevalence categories (e.g., SCC: n=30) resulted in wide confidence intervals, limiting statistical certainty. The retrospective design excluded clinical context (e.g., patient history), and ground-truth diagnoses combined histopathologically confirmed and unconfirmed (expert-annotated only) lesions, introducing potential verification bias. Finally, explainability analysis lacked standardized criteria for feature identification, possibly inflating perceived proficiency. These factors constrain the clinical applicability of our findings. To our knowledge, no prior studies have evaluated generic foundational GPT-4 Turbo Vision on a multiclass dermoscopic dataset, limiting direct comparison.

The study reveals a critical paradox in deploying general-purpose AI like GPT-4 Turbo Vision for dermatologic diagnosis: while its intuitive, no-preprocessing interface makes it dangerously easy for non-specialists to use, its subclinical accuracy (37.6% overall, with hazardous triage failures) renders it entirely unfit for real-world practice. The model's impressive generative fluency risks misleading clinicians, who may overestimate its reliability, especially as it frequently produces plausible-sounding feature descriptions despite

diagnostic inaccuracies. This gap between usability and performance highlights why these tools should, for now, remain strictly limited to research.

Although this study was not initially intended as a cautionary narrative, the findings revealed critical limitations that warrant careful consideration—particularly the model's subclinical accuracy and hazardous triage failures, which preclude clinical use. Expanded studies should include diverse populations, especially underrepresented skin types (Fitzpatrick IV–VI), and rare malignancies to assess diagnostic reliability in real-world settings. Prospective trials comparing the model with dermatologists in live clinical workflows are needed to evaluate its practical utility, while standardised explainability frameworks and bias mitigation strategies must be developed to address its diagnostic shortcomings. In light of these findings, the model is not yet ready for clinical trials. Instead, research should prioritise hybrid AI approaches to bridge the performance gap with specialised systems. Until such improvements are validated, GPT-4 Turbo Vision should remain confined to controlled research environments, with clear safeguards to prevent premature clinical adoption.

5 Ethical Considerations

This study uses publicly available, de-identified dermoscopic images from established repositories (ISIC Archive, DermNet NZ, dermatology textbooks), which are exempt from institutional review board approval. All images will be used in accordance with the terms of service and data usage agreements specified by the source repositories. If required by institutional policies or specific image sources, additional permissions will be obtained. The protocol acknowledges that demographic data (age, sex, ethnicity) are unavailable, making it impossible to assess discrepancies across sociodemographic groups. Explainability evaluations do not assess causal validity and may not reflect true model reasoning. No patients or public will be involved in any aspect of this study.

6 Funding

This research received no external or internal funding. No financial support was provided by any unding agency, institution, or commercial entity.

7 Conflicts of Interest

The authors declare that they have no conflicts of interest or financial disclosures that could be perceived to influence the study design, conduct, interpretation, or reporting of the research.

8 Registration

This study was pre-registered on the Open Science Framework (OSF) to enhance transparency and reduce potential biases in research design, data collection, and analysis. The full pre-registration protocol is publicly available and can be accessed using the following DOI: 10.17605/OSF.IO/HKMTV.

All hypotheses, methods, and planned analyses were specified in advance of data collection.

9 Study Data Availability

No patients or public will be involved in any aspect of this study. The study dataset (750 de-identified dermoscopic images with a detailed data dictionary) will be available from the authors upon request for review purposes and reproducibility, subject to compliance with the original repositories' usage agreements and citation requirements.

10 Code Sharing

No novel code was developed, as the study solely evaluates existing AI model using standard methods; hence, no unique code repository is available.

11 Statement on AI Assistance

In preparing the study protocol and manuscript text, the authors used ChatGPT-4 Turbo and DeepSeek-V3 (versions from March and June 2024) to assist with grammar refinement, syntax improvement, and idea generation. These tools were used solely for language and writing support, not for data analysis or model development. All AI-assisted content was thoroughly reviewed, edited, and verified by the authors, who assume full responsibility for the accuracy and integrity of the protocol.

12 Acknowledgements

I gratefully acknowledge the supervision and academic guidance of Ao.Univ.-Prof. Dr. med. univ. Harald Kittler, whose expertise and critical insights were essential to the successful completion of this thesis.

Images used in this study were sourced under Creative Commons licenses (Creative Commons Attribution, Creative Commons Attribution–ShareAlike & Creative Commons Zero):

- University of Athens, Andreas Syggros Hospital
- Memorial Sloan Kettering Cancer Center
- Hospital Clínic de Barcelona
- University of Queensland, Diamantina Institute & Dermatology Research Centre
- Sydney Melanoma Diagnostic Center, Royal Prince Alfred Hospital
- Hospital Italiano de Buenos Aires
- DermNet NZ

13 References:

1. Liu X, Duan C, Kim MK, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform. 2024;12:e59273. doi:10.2196/59273.

2. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. 2024;90(5):1057-1059. doi:10.1016/j.jaad.2023.12.062.

3. Wang J, Hu G. Boosting GPT-4V's accuracy in dermoscopic classification with few-shot learning. Comment on "Can ChatGPT vision diagnose melanoma?" J Am Acad Dermatol. 2024;91(6):e165-e166. doi:10.1016/j.jaad.2024.06.098.

4. Traini DO, Palmisano G, Peris K. Large language models and dermoscopy: assessing the potential of task-specific GPT-4 vision in diagnosing basal cell carcinoma. J Eur Acad Dermatol Venereol. 2024;38(12):2320-2322. doi:10.1111/jdv.20333.

5. Trager MH, Gordon ER, Breneman A, Kim E, Samie FH. Accuracy of ChatGPT in diagnosis and management of dermoscopic images. Arch Dermatol Res. 2025;317(1). doi:10.1007/s00403-024-03729-z.

6. Karampinis E, Toli O, Georgopoulou K-E, et al. Can artificial intelligence "hold" a dermoscope? Evaluation of an artificial intelligence chatbot to translate the dermoscopic language. Diagnostics (Basel). 2024;14(11):1165. doi:10.3390/diagnostics14111165.

7. Perlmutter JW, Milkovich J, Fremont S, Datta S, Mosa A. Beyond the surface: assessing GPT-4's accuracy in detecting melanoma and suspicious skin lesions from dermoscopic images. Plast Surg (Oakv). 2025;0(0). doi:10.1177/22925503251315489.

8. Duniphin D. Limited access to dermatology specialty care: barriers and teledermatology. Dermatol Pract Concept. 2023;13(1):e2023031. doi:10.5826/dpc.1301a31.

9. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799.

10. Collins GS, Moons KGM. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:q902. doi:10.1136/bmj.q902.

11. Zalaudek I, Piana S, Moscarella E, et al. Morphologic grading and treatment of facial actinic keratosis. Clin Dermatol. 2014;32(1):80-87. doi:10.1016/j.clindermatol.2013.05.028.
WHO Classification of Tumours Editorial Board. WHO Classification of Skin Tumours. 4th ed. Vol 11. Lyon, France: IARC Press; 2018:40-47.
12. Kim JYS, Kozlow JH, Mittal B, Moyer J, Olencki T, Rodgers P. Guidelines of care for the management of basal cell carcinoma. J Am Acad Dermatol. 2018;78(3):540-559. doi:10.1016/j.jaad.2017.10.006.
13. Keung EZ, Gershenwald JE. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. Expert Rev Anticancer Ther. 2018;18(8):775-784. doi:10.1080/14737140.2018.1489246.
14. Elder DE, Massi D, Scolyer RA, eds. WHO Classification of Skin Tumours. 4th ed. Lyon, France: International Agency for Research on Cancer; 2018.
15. Devi Vairamani A, Nayyar A. Prompt Engineering. Boca Raton, FL: CRC Press; 2024. doi:10.1201/9781032692319.
16. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193-204. doi:10.1016/j.jbi.2014.02.013.
17. Mirza FN, Lim RK, Yumeen S, et al. Performance of three large language models on dermatology board examinations. J Invest Dermatol. 2024;144(2):398-400. doi:10.1016/j.jid.2023.06.208.
19. Kittler H, Halpern A. How foundation models are shaking the foundation of medical knowledge. J Invest Dermatol. 2024;144(2):201-203. doi:10.1016/j.jid.2023.08.032.
20. Dinnes J, Deeks JJ, Chuchu N, et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database Syst Rev. 2018;12(12):CD011902. doi:10.1002/14651858.CD011902.pub2.
21. Menzies SW, Sinz C, Menzies M, et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. *Lancet Digit Health*. 2023;5(10):e679–e691. doi:10.1016/S2589-7500(23)00130-9
22. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2:665–673. doi:10.1038/s42256-020-00257-z.

23. Ong LC, Unnikrishnan B, Tadic T, et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. NPJ Digit Med. 2024;7:124. doi:10.1038/s41746-024-01118-4.
24. Sawyers EA, Wigle DT, Marghoob AA, Blum A. Dermoscopy training effect on diagnostic accuracy of skin lesions in Canadian family medicine physicians using the triage amalgamated dermoscopic algorithm. Dermatol Pract Concept. 2020;10(2):e2020035. doi:10.5826/dpc.1002a35.
25. Kittler H, Marghoob AA, Argenziano G, et al. Standardization of terminology in dermoscopy/dermatoscopy: results of the third consensus conference of the International Society of Dermoscopy. J Am Acad Dermatol. 2016;74(6):1093-1106. doi:10.1016/j.jaad.2015.12.038.