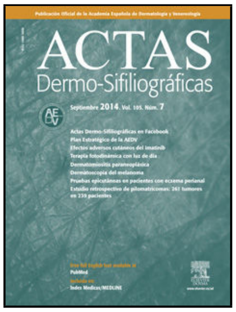


Foundation Models in Dermatology: Advances in Artificial Intelligence. A Narrative Review

D. Emilio Pimienta-Rosero E. Yesid Benavides-Tulcán DC. Fajardo-Murcia

PII: S0001-7310(25)00836-1
DOI: https://doi.org/doi:10.1016/j.ad.2025.104560
Reference: AD 104560

To appear in: *Actas dermosifiliograficas*

Received Date: 11 March 2025
Accepted Date: 28 April 2025

Please cite this article as: Emilio Pimienta-Rosero D, Yesid Benavides-Tulcán E, Fajardo-Murcia D, Foundation Models in Dermatology: Advances in Artificial Intelligence. A Narrative Review, *Actas dermosifiliograficas* (2025), doi: https://doi.org/10.1016/j.ad.2025.104560

Sección: **Review**

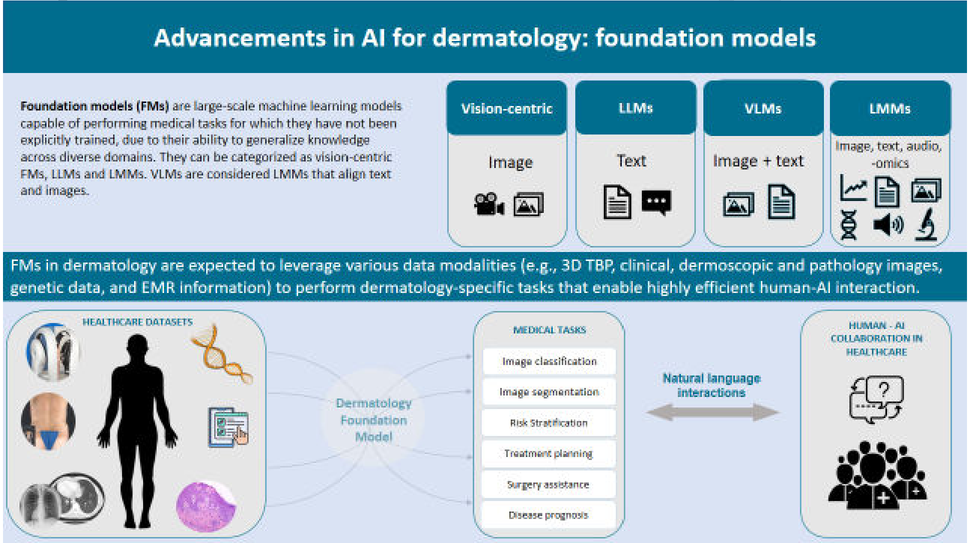
Foundation Models in Dermatology: Advances in Artificial Intelligence. A Narrative Review

D. Emilio Pimienta-Rosero¹,2; E. Yesid Benavides-Tulcán1,2; D. C. Fajardo-Murcia1,2,3. 1. Facultad de Salud. Departamento de medicina interna. Sección de dermatología. Universidad del Valle, Cali, Colombia. 2. Sección de dermatología. Hospital Universitario del Valle "Evaristo García", Cali, Colombia. 3. Dermatología estética, Universidad de Alcalá, España. Correspondencia Eine Yesid Benavides Tulcán eine.benavides@correounivalle.edu.co

Authors: C. Vico-Alonso, M. Sashindranath, S. Yan, Z. Yu, Z. Ge, and V. Mar

Corresponding author:
Cristina Vico-Alonso
E-mail address: c.vicoalonso@gmail.com

Graphical abstract



Advancements in AI for dermatology: foundation models

Foundation models (FMs) are large-scale machine learning models capable of performing medical tasks for which they have not been explicitly trained, due to their ability to generalize knowledge across diverse domains. They can be categorized as vision-centric FMs, LLMs and LMMs. VLMs are considered LMMs that align text and images.

FMs in dermatology are expected to leverage various data modalities (e.g., 3D TBP, clinical, dermoscopic and pathology images, genetic data, and EMR information) to perform dermatology-specific tasks that enable highly efficient human-AI interaction.

Abstract:
Foundation models (FMs) are deep learning models distinguished by their extensive training data and adaptability to a variety of medical tasks including disease classification, monitoring, risk stratification and treatment planning. They are categorized into large language models (LLMs) and vision-centric FMs, which process only text or only images as inputs, respectively. Alternatively, some models integrate multiple modalities, such as images and texts, in vision language models (VLMs), or may expand this multimodality by integrating audio, video, genomics and patient metadata in large multimodal models (LMMs). FMs are expected to help dermatologists in the clinical setting by leveraging advanced AI capabilities to standardize diagnosis and more accurately quantify disease severity, personalize treatment planning and ultimately improve patient outcomes. In this narrative review, we present an overview of the main milestones in generative AI that have driven the

evolution of dermatology-focused FMs, a field still under active research. Additionally, we summarize the current landscape and the principal medical FMs that have been developed for image-based medical specialties. Finally, we analyze potential risks and future directions in this field, offering insights from both clinical and technical perspectives.

Keywords: artificial intelligence, generative AI, foundation models, medical AI, LLMs, LMMs, VLMs, vision-centric FMs

INTRODUCTION

Traditional AI, that typically focuses on a single task (eg, classification or prediction) has been superseded by generative AI that can create outputs based on the features it has learned from a variety of training datasets, including images, text, video, audio and other sources of information[1]. Such capabilities are particularly attractive in medicine, where various domains of care (prevention, diagnosis, treatment, prognosis) rely on diverse information (photography, radiology imaging, video, genomics, medical textbook knowledge, etc). This new domain has led to the development of foundation models (FMs), a paradigm shift from limited-scale models in conventional deep learning algorithms, to large-scale models (Table 1).

A FM is a machine learning (ML) model that is trained on massive amounts of data which it can draw on to perform different tasks, including those for which it has not been explicitly trained[2]. By employing this training and adaptation methodology, FMs can significantly improve the performance of traditional AI models, leading to enhanced diagnostic accuracy, optimization of treatment recommendations, streamlined care and improved outcomes across multiple disciplines. The term "foundation model", first coined in 2021 by Bommasani, serves as an umbrella term encompassing various types of models. Vision-centric FMs focus purely on visual data (images or videos); large language models (LLMs) analyze text data while large multimodal models (LMMs) can be trained on multiple modalities. Thus, vision-language models (VLMs) that take both images and text as input are the paradigm example within this latest category[3]. Given its heavy reliance on visual diagnosis, dermatology is one of the primary medical specialties with immense potential for implementing these models.

The objective of this narrative review is to summarize the present state of FMs and current applications in dermatology. We also provide a dermatologist's perspective on the health care impact of these models on our field emphasizing the crucial need for fostering interdisciplinary collaborations across health care and engineering fields.

THE HISTORY OF FOUNDATION MODELS

The significant technological advancements in computerization between 2010 and 2016, coupled with increased access to large datasets, led the world to transition from traditional programming to ML (Figure 1). Unlike traditional programming, ML uses diverse algorithms that automatically generate rules based on the input data received, without prior programming. Initially, ML models extracted image features that were manually defined and annotated beforehand[4]. Subsequently, deep learning with convolutional neural networks (CNNs), rapidly gained interest in dermatology due to their high performance in extracting image features, classifying skin conditions, and differentiating malignancy from inflammatory diseases.[5-10] In dermatology AI studies, these networks based on publicly available CNN architectures, such as IncepcionV3, InceptionV4 and ResNet, are the most common architecture employed and they have outperformed dermatologists in controlled research environments[5].

In 2018, the adoption of transformers, a type of neural network architecture, for natural language processing (NLP) tasks, and later for image analysis, marked the inception of FMs[11]. Building on this foundation, vision transformers (ViT) revolutionized computer vision by leveraging attention mechanisms to analyze both text and image data in parallel, enabling faster and more efficient computations[12]. Due to their computational efficiency and scalability, these architectures provided capability to train large models with over 100 billion parameters. In the evolution of ML, FMs could be seen as a practically inevitable step, with the medical field emerging as a key adopter. AI is increasingly being used for image analysis in pathology, ophthalmology, and radiology, which have already developed and implemented medical FMs.[13-15]

In addition to the deep neural networks employed by FMs, methods used for training require a more complex computational design vs prior deep learning models. When developing an algorithm for a ML model, it can be trained in a supervised or unsupervised manner. Supervised learning has been widely used in AI models in dermatology so far, wherein the learning is guided by providing data that is labelled with a certain category. For example, clinical or dermoscopic images of cutaneous lesions are annotated with the confirmed histopathological diagnosis or expert consensus for unbiopsied lesions (eg, ground-truth). Unsupervised learning does not require labelled data but has not been common in dermatology research to date. However, FMs are often based on self-supervised learning (SSL). Considered a subtype of unsupervised learning, SSL models learn from data that has not been annotated. They learn to associate different words within a text in the case of LLMs; images and text in the case of VLMs; or they learn useful features and representations of the images in vision-centric FMs[15]. SSL is more scalable as it depends on unlabelled data and the models are forced to predict parts of the inputs, which makes them more useful than models trained on limited labels. This concept is also known as zero-shot learning, where the model can perform tasks for which it was not explicitly trained by understanding the task through NLP. Since there is a total absence of labelled examples, the model makes use of auxiliary information, such as attributes and semantic descriptions[16]. FMs apply SSL in the pre-training phase, wherein the model learns from large-scale unlabelled data to self-generated labels[4]. In a subsequent phase, the model acquires the ability to perform specific downstream tasks through a fine-tuning process or reinforcement learning, among others[2] (Figure 2). Fine-tuning is based on supervised learning, which requires only a small-scale annotated dataset, which means the model is tailored to specific tasks like summarization or question-answering. This reinforcement learning uses human feedback to improve outputs.[17,18] This innovative methodology is particularly compelling in dermatology, where large labelled imaging datasets are frequently unavailable, and human expert annotation can be time-consuming and prone to inaccuracy or misclassification.

TYPES OF FMs

Large language models (LLMs)

LLM have come into the spotlight in medical AI following their widespread availability to the public through platforms such as ChatGPT and BERT (Google). LLMs exhibit excellent performance in tasks related to NLP, such as translation, text generation or question-answering, demonstrating unprecedented ability to comprehend the intricacies of human language[19,20]. After the launch of the GPT series in 2019, the advancements in LLMs have immersed the world in global and direct human-machine interaction, particularly since the release of ChatGPT-3, which was pre-trained in 175 billion parameters. LLMs are pre-trained on extensive text data mainly extracted from online sources, including Wikipedia, Pubmed articles and electronic health records[2].

In 2023, Med-PaLM became the first general medical FM to pass the US Medical Licensing Examination (USMLE), achieving a score of 67.6% on the MedQA dataset, a comprehensive dataset derived from professional medical board exams. However, the accuracy remained inferior to clinicians[20-22]. Shortly after, with the release of MedPaLM-2, this score was surpassed reaching 86.5%, which could be considered expert level[23]. On the GPT-series, the performance of LLMs on USMLE exams was evaluated obtaining scores around 60%, which falls at or near the passing threshold for all required exams[24]. In late 2023, the performance of LLMs on the MedQA dataset experienced a significant milestone when ChatGPT4-Medprompt achieved a 90.2% accuracy rate, the highest score to date, outperforming the then state-of-the-art Med-PaLM 2[25]. Notably, in this instance, the authors did not employ extensive fine-tuning, opening new opportunities in FMs for medicine. However, it should be noted that all these models were deployed in specific research scenarios and have not yet been tested in real-world contexts.

In dermatology, they have also been shown to aid in passing dermatology certification exams, with an overall accuracy of 90% for ChatGPT-4 in passing the specialty examination, although the authors emphasized the possibility that the answers could have been used as part of the training dataset[26]. Language seems to play a major role for model performance. Studies have shown variations in performance on dermatology certificate exams based on the language used to answer. While ChatGPT-4 would pass the test regardless of the language utilized, ChatGPT-3.5 failed an exam conducted in Polish[27]. The study by Mirza et al. assessed the performance of Google Bard, ChatGPT-3.5 and ChatGPT-4 on mock dermatology board exams[28]. Not surprisingly, ChatGPT-4 achieved the highest overall performance among the 3 LLMs. Nonetheless, it was notably the sole model to demonstrate accuracy associated with both adjusted and unadjusted readability levels. In contrast, another study reported that Google Bard outperformed ChatGPT-3.5 in overall accuracy, although the authors noted important limitations, such as not including image-based questions and failing to compare the models' performance with that of dermatology residents. These shortcomings may reduce the applicability of the results in real-world clinical or exam scenarios[29].

Researchers have investigated the potential applications of these models in dermatology consultations, including assistance with triaging GP referrals to dermatology services[30]. In this proof-of-concept study, despite the small dataset size (268 referral letters), the authors showed that BERT could help categorise patients into binary outcomes (routine vs non-routine cases).

From a health care provider's perspective, Jin et al. suggest various representative prompts to interpret laboratory findings and manage overall skin conditions[31]. Moreover, they describe how ChatGPT could support patients' education across multiple dermatological domains, aid medical students as a learning tool, and facilitate triage systems in clinical trials by assessing eligibility criteria. Other authors have highlighted additional applications within dermatology, such as assisting in writing clinical guidelines, helping draft academic leaflets from academic societies and providing letters to patients[32]. Interestingly, a pilot study has shown that after-visit summaries generated by ChatGPT-3.5 in dermatology consultations achieved high patient satisfaction scores[33].

Other studies have also explored the utility of LLMs for patient queries about melanoma. A study evaluating the utility of ChatGPT's responses found them to be accurate, with a mean score of 4.88 out of 5, rated by three board-certified dermatologists. However, while the content was precise, the text generated might not be suitable for the general public due to its advanced level[34].

Furthermore, the performance of LLMs has been investigated in Mohs surgery. A study showed that ChatGPT significantly improved frequent patient queries regarding this type of surgery, although 2 responses were considered harmful and inaccurate[35]. Compared with what has been previously published, a recent study comparing ChatGPT and Google Bard, found ChatGPT to be comparable to surgeons answers in terms of accuracy and comprehensibility. They attribute this variability in performance to the absence of restrictions imposed on the chatbot to emulate patient querying style[36]. The accuracy between dermatology surgeons and ChatGPT has also been tested to select reconstruction techniques for Mohs surgery defects, showing substantial variability between surgeons and chatbot choices and raising concern about the inadvertent reliance on these chatbots[37].

Lately, the scope of LLMs have extended into inflammatory diseases in dermatology. In acne and atopic dermatitis (AD), ChatGPT-3.5 achieved an average readability FRES score (Flesch Reading Ease Score) of 39.88 for acne and 30.13 for AD queries, which corresponds to college-level comprehension[38]. Interestingly, some knowledge gaps were found, particularly regarding available drugs such as spironolactone for acne treatment or biologics and JAK inhibitors for the management of AD. In psoriasis, LLMs have been reportedly comparable to meta-analyses in generating conclusions assessing drug efficacy. Yet, ChatGPT remains insufficient when analyzing more than 3 drugs[39].

Large multimodal models (LMMs)

Due to the visual precision required in dermatology, LMMs that comprehend and generate new content based on the alignment of different sources are particularly attractive as a promising solution to the scarcity of labelled data in medicine[40]. In the clinical setting, these LMMs could be used across various tasks that mirror the real-world practices in dermatology, wherein an accurate clinical diagnosis is supported by visual inputs (dermoscopy, confocal microscopy, ultrasonography, pathology); attributes intrinsic of the individual (phenotype, past medical history, etc); and genetic and laboratory findings (proteomics, genomics, transcriptomics).

LMMs in medicine are currently a promising approach to exploring the performance of various sources of data on specific medicine tasks with an enormous potential[41]. In fact, multimodal learning has been reportedly proven to outperform single-source models[42]. The ultimate definition of a multimodal model involves using and understanding information from multiple modalities. Therefore, VLMs can be considered multimodal systems due to their ability to process large-scale language and visual data. The fusion of visual and textual inputs through sophisticated AI architectures forms the core of VLMs, enabling the generation of text-to-image, visual-question answering or image-captioning[43] (Figure 3).

In 2021, the first state-of-art technology CLIP (contrastive language-image pretraining) was able to match images with their most relevant descriptions, and vice versa. Through a contrastive learning approach, a form of SSL, the model can effectively process both images and text with promising results across several tasks due to its generalizability and interpretability[44]. The use of CLIP in the medical field has recently been explored leveraging multimodal medical imaging like X-rays, MRI, and CT scans with potential implications that still need to be fully addressed and contextualised[45]. In the field of dermatological research, several studies have explored multimodality. Clinical, dermoscopic and fluorescence images have been combined with patient metadata in small-scale skin cancer research studies, with sex, age and lesion location being the most utilized clinical inputs[46-48]. Diagnostic accuracy has been reported to benefit from this multimodal integration[48]. On the contrary, it has been shown that the benefit of integrating patient metadata into some AI models might be limited[49]. Yet, large-scale multimodal models in dermatology are still scarce[50].

Another medical VLM, DALL.E2, addressed the limitations of biassed training datasets in dermatology. Synthetic data of skin lesions generated by models like DALL.E2 can effectively augment the training datasets to improve the performance of a conventional dermatological classifier[51]. The shortage of labelled health care data is prompting a rise of synthetic data in ML for medicine[52]. Interestingly, some authors have demonstrated that synthetic data cannot surpass real-world data diversity in improving medical AI algorithms[53]. To help generate accurate labels, a FM has been recently released to annotate skin lesions images with dermatology terms that can be used to train algorithms[54].

Given the crucial role of histopathology in precise dermatological diagnosis and the appearance of digital pathology through whole slide images, VLMs in this field are also steadily increasing. A VLM was trained using dermatopathology images and corresponding labels from medical Twitter, demonstrating impressive performance, and a recent FM in histopathology has been pre-trained using 1.17 million images with the corresponding text description[55].

In a recent study, the accuracy of ChatGPT vision in diagnosing melanomas was evaluated using dermoscopic images from the ISIC archive. The model showed a sensitivity of 32% and specificity of 40% when only the first diagnosis suggested by the chatbot was considered[56]. It should be highlighted that this accuracy still falls behind that of publicly available diagnostic apps[57].

A new interactive VLM, called SkinGPT4, has been developed by fine-tuning MiniGPT-4 and trained on a vast collection of 52,929 skin disease images and clinician reports[58]. Consumers can upload their own skin photographs for analysis and categorization into specific groups, providing interactive treatment planning. However, the model has only been tested in 150 real-life cases, achieving a 78.76% accuracy in correct or relevant diagnosis and 83.13% accuracy in treatment recommendations upon evaluation of chatbot responses by certified dermatologists.

Vision-centric FMs

Despite the promise that multimodal learning holds in dermatology, they are still rare. This scarcity stems from the lack of large-scale, high-quality multimodal medical datasets, making their development extraordinarily challenging. Additionally, currently, these models require significant computing power, which most centers do not have access to. Currently, AI research in dermatology primarily focuses on vision-centric FMs, specialized in understanding visual information solely, which have now become a reference in image analysis for dermatology.

Since diagnosing skin conditions requires a full domain of different imaging modalities, these models aim to emulate dermatologists' visual expertise through the integration of skin images to adapt afterwards to various applications, including change detection, risk assessment, or lesion segmentation.

In 2022, a vision-centric FM for medical images across different image-based medical specialties was released[59]. However, this model was still limited to a single task in dermatology, which was image classification of 26 skin conditions. Another vision-centric FM in dermatology has been pre-trained using 1.18 million unannotated and coarsely labelled dermatology images collected from online sources and dermatologists' personal channels in China[60]. The model can categorize up to 22 skin diseases, including skin cancer, inflammatory and pigmented disorders. The model achieved an accuracy of 49.64% and has been made accessible to doctors through the launch of a WeChat-based app.

Yan et al. have recently introduced the first general-purpose multimodal FM specifically designed for dermatology. The model was able to classify up to 74 skin conditions and performs well on a range of interconnected clinical tasks such as skin cancer diagnosis, lesion monitoring, risk assessment, and predicting melanoma metastasis across different modalities. Interestingly, the model outperformed the average dermatologist by 10%, demonstrating superior ability in detecting early-stage melanoma[61]. Although these vision-based FMs show great promise as assistants not only in dermatology but also for general practitioners, they must first be analyzed as potential human-AI collaborators and undergo final validation in real-world clinical settings.

Table 2 illustrates the principal medical FMs to date, some of which include tasks in the field of dermatology.

RISKS AND CONCERNS

The application of FMs in medicine holds an unparalleled transformation in global health care delivery. However, they remain a subject of active research in dermatology across leading institutions worldwide. While some hospitals have steadily increased their presence within clinical practice (eg, AI scribes for patient consultations, triage tools for referrals), large-scale AI-driven diagnostic and management tools have yet to be integrated into a real health care scenario[62]. Additionally, several challenges and ethical considerations for the use of this technology in medicine need to be addressed.

A primary concern is the reliability, applicability and generalization of FMs, that are significantly impacted by the quality and quantity of the training dataset. Despite notable enhancement in overall performance for LLMs in medical questions, they are not exempt from errors. Concerns arise regarding potential bias in training data, where models may have inadvertently learned from the same questions they are later queried on to evaluate their performance. Although bias in training datasets has been a matter of discussion for decades, advances in neural network architectures have now made it possible to identify biases inherent in the data[63,64]. Researchers have shown that modern AI models can uncover hidden bias in datasets, particularly in large-scale datasets that are often less curated, and presumably less biassed[64]. Dermatological algorithms to classify cutaneous diseases may perform poorly on underrepresented skin types, sex, language and culture[65]. Since these models are predominantly trained in English, there exists an inherent bias against non-English-speaking regions. Addressing these biases is crucial, and seeking international collaboration on diverse, open-access datasets is essential. It is mandatory to break down international barriers and promote data sharing between countries to enhance the quality of datasets used in medical AI research, ultimately improving data inclusivity.

Secondly, once deployed, these models quickly become obsolete as their underlying training data evolves rapidly, potentially necessitating swift changes in prior established medical paradigm, such as new drugs discoveries, clinical guidelines, or staging systems. One significant barrier to real-time implementation of FMs is the substantial cost and magnitude of computing power required. Given their large-scale amount of data needs during the pre-training phase, these models are not equally accessible and may find it difficult to reach the entire population equally. Moreover, these highly energy-dependent models require significant computational resources, such as modern multi-core computers with powerful GPU cores (eg, NVIDIA SuperPod) in massive data centres[2,66].

Publicly available AI-systems capable of generating rapid responses within seconds may alter the medical principles of nonmaleficence and justice[67]. On one hand, if AI is entitled to make clinical decisions, errors may occur, raising uncertainties about liability for diagnosis and treatment recommendations made by AI models[67,68]. The generation of medically inaccurate or fabricated content ("hallucinations") also poses a significant challenge to the safe and ethical use of these systems[90]. On the other hand, the potential integration of these technologies into care systems raises concerns about equitable access, particularly among lay-persons who may face barriers related to financial resources or geographical location[65].

Experts emphasize that FMs should assist, not replace, health care providers[69]. They would work collaboratively with clinicians to provide solutions for complex clinical workflows (bedside decision support for inpatients, treatment comparison, surgical margins assessment, etc) and act as a triage system for non-urgent skin conditions[70]. Some authors have advocated for the explicit acquisition of informed consent before using these models, ensuring adherence to ethical principles of autonomy and transparency[91]. Additionally, FMs could establish a robust knowledge base, while health care providers would contribute the essential human element of the patient-doctor relationship, a critical aspect that has been significantly affected by the growing digital workload in recent decades.

Finally, for FMs to gain full acceptance, they must undergo rigorous regulatory scrutiny, akin to medical devices to ensure confidentiality and security[71]. Since the core of these models is clinical data, potential security breaches may occur, and it should be noted that for most models patients may not have explicitly provided consent for their data use[72-75]. In this regard, the research community should raise for discussion the implementation of robust privacy policies within the health care system to explicitly address data use for AI. This includes establishing frameworks to obtain informed consent, ensuring transparency in data handling, and protecting patient information while responsibly collecting data.

CONCLUSIONS AND FUTURE PERSPECTIVES

FMs in dermatology have immense potential for application in primary care triaging and holistic disease management. Their adoption could help alleviate the shortage of dermatologists in certain regions and improve health care access for under-resourced and geographically isolated populations[76]. Additionally, they may discover clinical associations and patterns that enhance disease management, reducing the enormous effort and cost of clinical trials to tailor treatments more effectively to individual patients. A new era in drug discovery for dermatology could unfold, particularly after the release of models like AlphaFold that could contribute to this area of dermatology research[77]. However, while there is still a considerable journey ahead for AI to advance sufficiently to handle complex medical scenarios in dermatology, the development of large-scale, high-quality multimodal datasets is recognized as essential and inevitable to achieving these goals, despite potential limitations related to privacy and ethics in accessing extensive datasets [78,79].

The application of FMs is poised to fulfil the aspiration of personalized and individualized medicine, particularly in specialties whose knowledge core is inherently visual. Nevertheless, skin specialists should embrace rather than fear the implementation of this technology, actively participating in its development within healthcare. To facilitate this transition, dermatologists must receive proper training in AI and FMs, providing them with the necessary tools to understand how these algorithms function and how to identify potential biases that can affect their performance. Furthermore, the effectiveness of human-AI collaboration should be evaluated, with a focus on how health care professionals interact with AI-generated recommendations. Behavioural differences among doctors must also be considered. The level of seniority, along with personal traits like confidence or insecurity, can condition a doctor’s attitude toward AI suggestions. Addressing these factors will lead to an effective collaboration with these AI models to achieve a more interactive, multimodal, and integrated health care delivery system.

REFERENCES

1. Gui H, Omiye JA, Chang CT, Daneshjou R. The Promises and Perils of Foundation Models in Dermatology. J Invest Dermatol. 2024;144:1440-8.
2. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:210807258. 2021.
3. Omiye JA, Gui H, Daneshjou R, et al. Principles, applications, and future of artificial intelligence in dermatology. Front Med (Lausanne). 2023;10:1278232.
4. Schneider J, Meske C, Kuss P. Foundation models: a new paradigm for artificial intelligence. Bus Inf Syst Eng. 2024:1-11.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-8.
6. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 2018;138:1529-38.
7. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer. 2019;111:30-7.
8. Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol. 2019;155:58-65.
9. Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol. 2020;156:29-37.
10. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26:900–908.
11. Parmar N, Vaswani A, Uszkoreit J, et al., editors. Image transformer. In: International conference on machine learning; 2018. PMLR.
12. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. 2020. Available at: arXiv:2010.11929.
13. Huang Z, Bianchi F, Yuksekgonul M, et al. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307-16.
14. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622:156-63.
15. Pai S, Bontempi D, Hadzic I, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell. 2024;6(3):354-67.
16. Sathya R, Abraham A. Comparison of supervised and unsupervised learning algorithms for pattern classification. Int J Adv Res Artif Intell. 2013;2:34-8.
17. Mall U, Hariharan B, Bala K, editors. Field-guide-inspired zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.
18. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730-44
19. Liu J, Yang M, Yu Y, et al. Large language models in bioinformatics: applications and perspectives. arXiv. 2024.
20. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80.
21. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. NEJM AI. 2024;1(3).
22. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11:6421.

23, Singhal K, Tu T, Gottweis J, et al. (2025). Toward expert-level medical question answering with large language models. Nat Med. 2025;31:943–950. https://doi.org/10.1038/s41591-024-03423-7.
24. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2).
25. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv preprint. 2023. Available at: arXiv:2311.16452.
26. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clin Exp Dermatol. 2024;49:722-7.
27. Lewandowski M, Łukowicz P, Świetlik D, et al. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol. 2024;49:686-91.
28. Mirza FN, Lim RK, Yumeen S, et al. Performance of three large language models on dermatology board examinations. J Invest Dermatol. 2024;144:398-400.
29. Park L, Ehlert B, Susla L, et al. Performance of large language model artificial intelligence on dermatology board exam questions. Clin Exp Dermatol. 2024; 25:49:733-734. doi: 10.1093/ced/llad355.
30. Wang M, Kloczko E, Altayeb A, et al. Towards automated dermatology triage: deep learning and knowledge-driven approaches. Available at: SSRN 4385662. 2023.
31. Jin JQ, Dobry AS. ChatGPT for health care providers and patients: Practical implications within dermatology. J Am Acad Dermatol. 2023;89:870-1.
32. Kluger N. Potential applications of ChatGPT in dermatology. J Eur Acad Dermatol Venereol. 2023;37:e941-e2.
33. Young AT, Lane BN, Ozog D, Matthews NH. Patients and dermatologists are largely satisfied with ChatGPT-generated after-visit summaries: A pilot study. JAAD Int. 2024;15:33-5.
34. Young JN, Ross OH, Poplausky D, et al. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. J Am Acad Dermatol. 2023;89:602-4.
35. O'Hern K, Rames MM, Rames JD, et al. ChatGPT improves readability of clinical responses to questions about Mohs surgery but may misinform. Dermatol Surg. 2022. doi:10.1097.
36. Robinson MA, Belzberg M, Thakker S, et al. Assessing the accuracy, usefulness, and readability of artificial-intelligence-generated responses to common dermatologic surgery questions for patient education: A double-blinded comparative study of ChatGPT and Google Bard. J Am Acad Dermatol. 2024;90:1078-80.
37. Cuellar-Barboza A, Brussolo-Marroquín E, Cordero-Martinez FC, et al. An evaluation of 'ChatGPT' compared to dermatological surgeons' choice of reconstruction of Mohs surgical defects. Clin Exp Dermatol. 2024.
38. Lau CB, Lilly E, Yu J, et al. Evaluating the efficacy of ChatGPT in addressing patient queries about acne and atopic dermatitis. Clin Exp Dermatol. 2024.
39. Lam Hoai XL, Simonart T. Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in Moderate-to-Severe Plaque Psoriasis. J Clin Med. 2023.
40. Shrestha P, Amgain S, Khanal B, et al. Medical vision language pretraining: A survey. *arXiv preprint*. 2023. Available at: arXiv:2312.06224.
41. Qiu J, Li L, Sun J, et al. Large AI models in health informatics: Applications, challenges, and the future. IEEE J Biomed Health Inform. 2023.

42. Huang Y, Du C, Xue Z, et al. What makes multi-modal learning better than single (provably). Adv Neural Inf Process Syst. 2021;34:10944-56.
43. Zhou K, Yang J, Loy CC, et al. Learning to prompt for vision-language models. Int J Comput Vis. 2022;130:2337-48.
44. Radford A, Kim JW, Hallacy C, et al., editors. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning; 2021: PMLR.
45. Zhao Z, Liu Y, Wu H, et al. CLIP in medical imaging: A survey. Med Image Anal. 202; 102, 103551. https://doi.org/10.1016/j.media.2025.103551.
46. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. Exp Dermatol. 2018;27:1261-7.
47. Binder M, Kittler H, Dreiseitl S, et al. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. Melanoma Res. 2000;10(6):556-61.
48. Höhn J, Hekler A, Krieghoff-Henning E, et al. Integrating patient data into skin cancer classification using convolutional neural networks: Systematic review. J Med Internet Res. 2021.
49. Höhn J, Krieghoff-Henning E, Jutzi TB, et al. Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification. Eur J Cancer. 2021;149:94-101.
50. Luo N, Zhong X, Su L, et al. Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal. Comput Biol Med. 2023;165:107413.
51. Sagers LW, Diao JA, Groh M, et al. Improving dermatology classifiers across populations using images generated by large diffusion models. arXiv preprint. 2022. Available at: arXiv:2211.13352.
52. Chen RJ, Lu MY, Chen TY, et al. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493-7.
53. Sagers LW, Diao JA, Melas-Kyriazi L, et al. Augmenting medical image classifiers with synthetic data from latent diffusion models. arXiv preprint. 2023. Available at: arXiv:2308.12453.
54. Kim C, Gadgil SU, DeGrave AJ, et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. Nat Med. 2024;30:1154-1165.
55. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30:863-74.
56. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. 2024;90:1057-9.
57. Sun MD, Kentley J, Mehta P, et al. Accuracy of commercially available smartphone applications for the detection of melanoma. Br J Dermatol. 2022;186:744-6.
58. Zhou J, He X, Sun L, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun. 2024;15:5649.
59. Azizi S, Culp L, Freyberg J, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat Biomed Eng. 2023;7:756-79.
60. Shen Y, Li H, Sun C, et al. Optimizing skin disease diagnosis: harnessing online community data with contrastive learning and clustering techniques. NPJ Digit Med. 2024;7:28.
61. Yan, S, Yu Z, Primiero C. et al. A multimodal vision foundation model for clinical dermatology. Nat Med. 2025. https://doi.org/10.1038/s41591-025-03747-y.

62. Gupta AK, Talukder M, Wang T, et al. The Arrival of Artificial Intelligence Large Language Models and Vision-Language Models: a potential to possible change in the paradigm of health care delivery in dermatology. J Invest Dermatol. 2024;144:1186-8.
63. Torralba A, Efros AA, editors. Unbiased look at dataset bias. *CVPR* 2011; 2011: IEEE.
64. Liu Z, He K. A decade's battle on dataset bias: are we there yet? arXiv preprint arXiv:240308632. 2024.
65. Li H, Moon JT, Purkayastha S, et al. Ethics of large language models in medicine and medical research. Lancet Digit Health. 2023;5(6).
66. Yuan Y. On the power of foundation models. International Conference on Machine Learning; 2023: PMLR.
67. Gordon ER, Trager MH, Kontos D, et al. Ethical considerations for artificial intelligence in dermatology: a scoping review. Br J Dermatol. 2024.
68. Lim BCW, Flaherty G. Artificial intelligence in dermatology: are we there yet? Br J Dermatol. 2019;181:190-1.
69. Kittler H, Halpern A. How foundation models are shaking the foundation of medical knowledge. J Invest Dermatol. 2024;144:201-3.
70. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259-65.
71. Malvehy J, Ginsberg R, Sampietro-Colom L, et al. New regulation of medical devices in the EU: impact in dermatology. J Eur Acad Dermatol Venereol. 2022;36:360-4.
72. Hogarty DT, Su JC, Phan K, et al. Artificial intelligence in dermatology—where we are and the way to the future: a review. Am J Clin Dermatol. 2020;21:41-7.
73. Rundle CW, Hollingsworth P, Dellavalle RP. Artificial intelligence in dermatology. Clin Dermatol. 2021;39:657-66.
74. Daneshjou R, Smith MP, Sun MD, et al. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol. 2021;157(11):1362-9.
75. Ferreira AL, Lipoff JB. The complex ethics of applying ChatGPT and language model artificial intelligence in dermatology. J Am Acad Dermatol. 2023;89(4).
76. Xu S, Gui H, Rotemberg V, et al. A framework for evaluating the efficacy of foundation embedding models in healthcare. *medRxiv*. 2024:2024.04.17.24305983.
77. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024:1-3.
78. HE, Yuting, et al. Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Rev Biomed Eng. 2025;18:172-191. doi: 10.1109/RBME.2024.3496744.
79. Wu J, Liu X, Li M, et al. Clinical text datasets for medical artificial intelligence and large language models—a systematic review. NEJM AI. 2024;1(6).
80. Tiu E, Talius E, Patel P, et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng. 2022;6:1399-406.
81. Christensen M, Vukadinovic M, Yuan N, et al. Vision–language foundation model for echocardiogram interpretation. Nat Med. 2024:1-8.
82. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30:850-62.
83. Lu MY, Chen B, Williamson DF, et al. A multimodal generative AI copilot for human pathology. Nature. 2024:1-3.
84. Vorontsov E, Bozkurt A, Casson A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med. 2024:1-12.
85. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024:1-8.

86. Zhang K, Zhou R, Adhikarla E, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024:1-13.
87. Zhang S, Xu Y, Usuyama N, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI. 2025; 2, AIoa2400640.
88. Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data. Proceedings of the 24th International Conference on Machine Learning; 2007.
89. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.
90. Goktas P, Grzybowski A. Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review. J Clin Med. 2024:3;13:5909. doi: 10.3390/jcm13195909. PMID: 39407969.
91. Gordon ER, Trager MH, Breneman A, et al. Chatting ethically: practical recommendations for ethical use of large language models in dermatology practice, research and education. Clin Exp Dermatol. 2024;50:175-176. doi: 10.1093/ced/llae335.

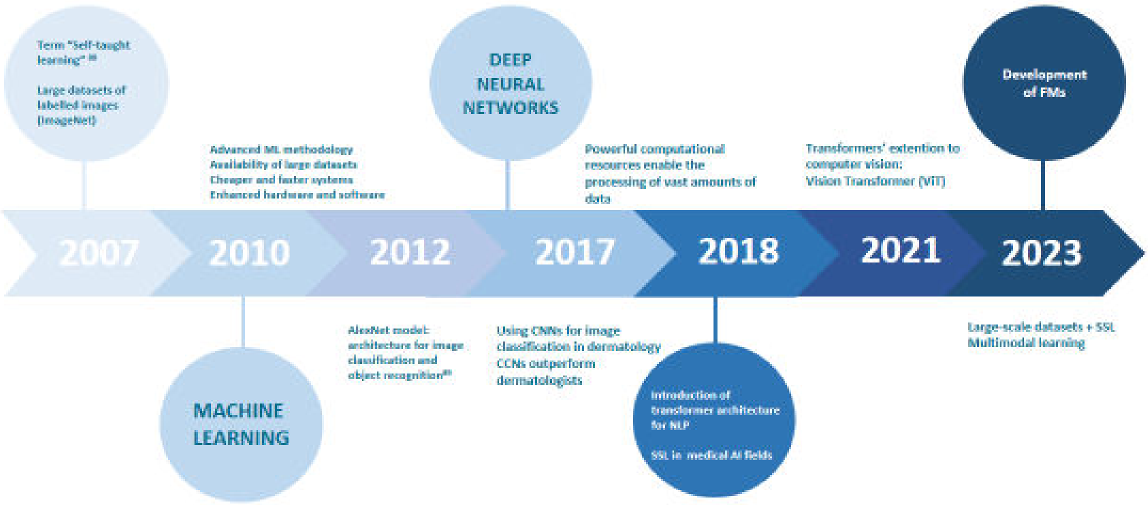

Term "Self-taught learning"
Large datasets of labelled images (ImageNet)
Advanced ML methodology
Availability of large datasets
Cheaper and faster systems
Enhanced hardware and software
DEEP NEURAL NETWORKS
Powerful computational resources enable the processing of vast amounts of data
Transformers' extention to computer vision: Vision Transformer (ViT)
Development of FMs
2007
2010
2012
2017
2018
2021
2023
MACHINE LEARNING
AlexNet model: architecture for image classification and object recognition
Using CNNs for image classification in dermatology
CCNs outperform dermatologists
Introduction of transformer architecture for NLP
SSL in medical AI fields
Large-scale datasets + SSL
Multimodal learning

Figure 1. Timeline of key milestones in the development of foundation models (FMs) from 2007 through 2023. ML: machine learning. CNN: convolutional neural networks. SSL: self-supervized learning.

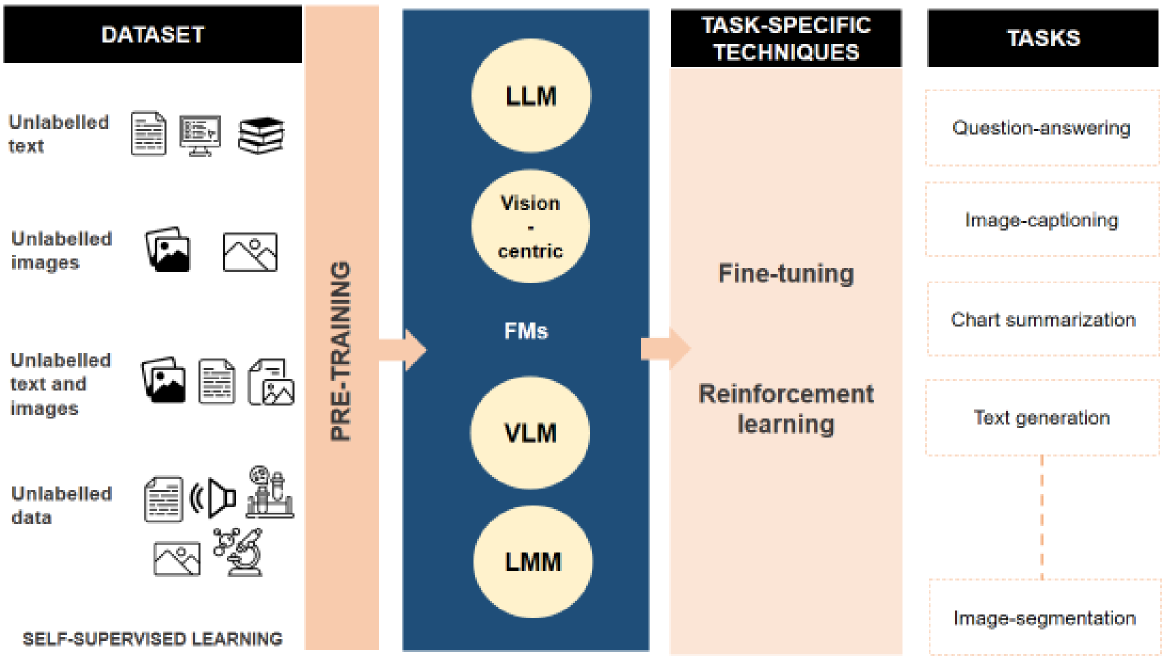


Figure 2. Foundation models (FMs) are pre-trained on diverse and large-scale datasets that include various modalities, often without labels. This pre-training phase enables the model to learn general representations and patterns. In a subsequent phase, the pre-trained model can be adapted to a wide range of specific tasks through fine-tuning or reinforcement learning, making it highly versatile and capable of performing numerous tasks.

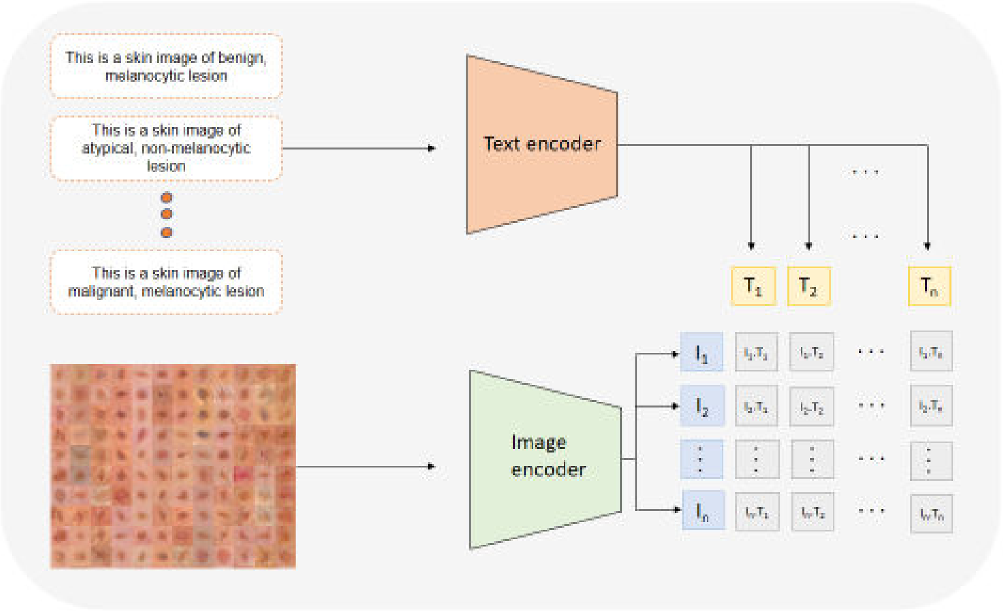


Figure 3. This figure depicts the vision-language pre-training phase for skin lesion classification. Text descriptions of skin lesions are processed by a text encoder, while a skin lesion dataset is concurrently and independently analyzed by an image encoder. The text encoder transforms the textual descriptions into a series of text embeddings in the token format (T1, T2,..., Tn), and the image encoder converts images into a series of image embeddings (I1, I2,…, In). These embedding tokens are then combined, forming joint representations that can be used for subsequent downstream tasks.