SkinGPT: A Dermatology Diagnostic System with Vision Large Language Model

Juexiao Zhou[1,2], Xin Gao[1,2,*]

Abstract—Skin and subcutaneous diseases are among the major causes of the nonfatal disease burden worldwide, affecting a significant proportion of the population. However, there are three major challenges in the field of dermatology diagnosis. Firstly, there is a shortage of dermatologists available to diagnose patients. Secondly, accurately diagnosing dermatological pictures can be challenging. Lastly, providing user-friendly diagnostic reports can be difficult. Recent advancements in the field of large language models (LLMs) have shown potential for clinical applications. However, current LLMs have difficulty processing images, and there are potential privacy concerns associated with using ChatGPT's API for uploading data. In this paper, we propose SkinGPT, which is the first dermatology diagnostic system that utilizes an advanced vision-based large language model. SkinGPT is the first system of its kind, incorporating a fine-tuned version of MiniGPT-4 with a vast collection of in-house skin disease images, accompanied by doctor's notes. With SkinGPT, users can upload their own skin photos for diagnosis, and the system can autonomously determine the characteristics and categories of skin conditions, perform analysis, and provide treatment recommendations. The ability to deploy it locally and protect user privacy makes SkinGPT an attractive option for patients seeking an accurate and reliable diagnosis of their skin conditions.

Index Terms—Dermatology, Deep learning, Large language model

1 Introduction

Skin and subcutaneous diseases rank as the fourth major cause of nonfatal disease burden worldwide, affecting a considerable proportion of individuals, with a prevalence ranging from 30% to 70% across all ages and regions [1]. However, dermatologists are consistently in short supply, particularly in rural areas, and consultation costs are on the rise [2], [3], [4]. As a result, the responsibility of diagnosis often falls on non-specialists such as primary care physicians, nurse practitioners, and physician assistants, which may have limited knowledge and training [5] and low accuracy on diagnosis [6], [7]. The use of store-and-forward teledermatology has become dramatically popular in order to expand the range of services available to medical professionals [8], which involves transmitting digital images of the affected skin area (usually taken using a digital camera or smartphone) [9] and other relevant medical information from users to dermatologists. Then, the dermatologist reviews the case remotely and advises on diagnosis, workup, treatment and follow-up recommendations [10], [11]. However, three main challenges remain, the first is the shortage of dermatologists [12], the second is the accurate diagnosis of dermatological pictures, and the third is the provision of user-friendly diagnostic reports [4], [13].

Advancements in technology have led to the development of various tools and techniques to aid dermatologists in their diagnosis [13]. For example, the development of artificial intelligence tools to aid in the diagnosis of skin disorders from images has been made possible by recent advancements in deep learning. Most studies have predominantly concentrated on identifying skin lesions through dermoscopic images [14], [15], [16]. However, dermatoscopy is often not readily available outside of dermatology clinics. Some studies have explored the use of clinical photographs of skin cancer [17], onychomycosis [18], and skin lesions on educational websites [19]. However, most of them focus on a single diagnosis rather than a comprehensive differential diagnosis, which is a scale used to guide treatment in cases of the uncertainty of diagnosis. As the state-of-the-art method, Yuan Liu et al. [4] presented a deep learning system for identifying the 26 most common dermatological conditions in adult cases that are referred for teledermatological consultation. Nevertheless, their approach still requires further analysis by physicians to issue reports and make clinical decisions and is unable to automatically generate detailed reports in text format and allow interactive dialogues. At present, there are no diagnostic systems available for users

[1]*Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia*
[2]*Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia*
[*]*Corresponding author. e-mail: xin.gao@kaust.edu.sa*

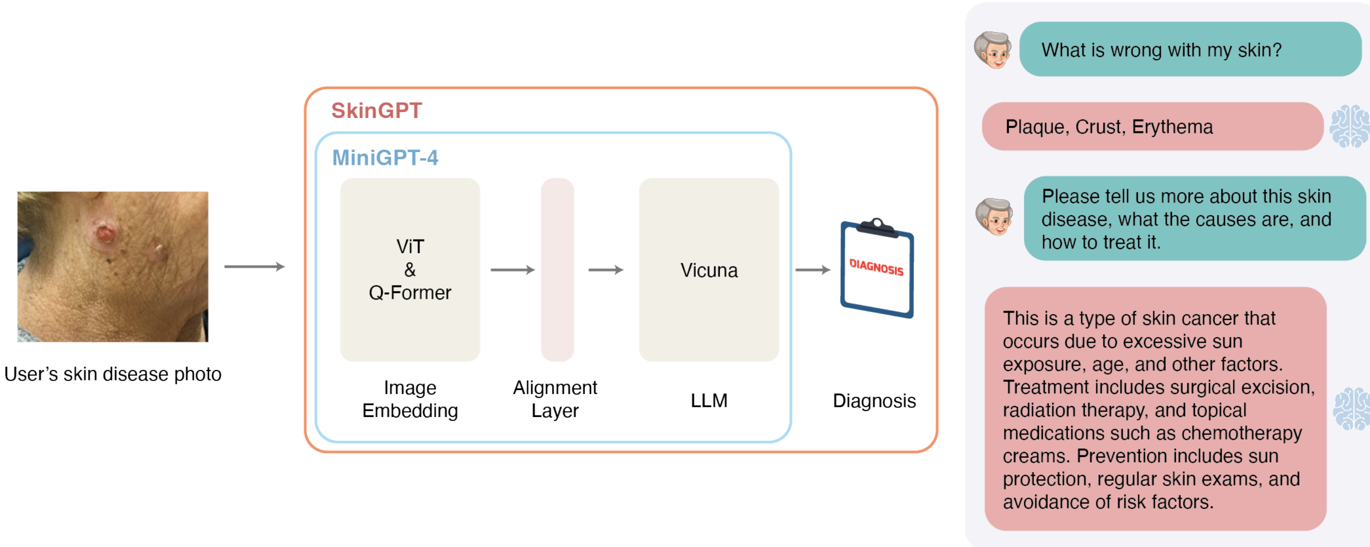

Fig. 1. Illustration of SkinGPT. SkinGPT is fine-tuned on MiniGPT-4 using large in-house skin disease images along with doctor's notes to allow for medical diagnosis and interactive dialogue. On the right is an example of interactive dialogue.

to self-diagnose skin conditions by submitting images that can automatically analyze and generate easy-to-understand text reports.

Over the past few months, the field of large language models (LLMs) has seen significant advancements [20], offering remarkable language comprehension abilities and the potential to perform complex linguistic tasks. One of the most anticipated models is GPT-4 [21], a large-scale multimodal model that has demonstrated exceptional capabilities, such as generating accurate and detailed image descriptions, providing explanations for atypical visual occurrences, and even constructing websites based on handwritten textual descriptions. Despite these remarkable advancements, GPT-4 is still not accessible to the public and is closed-source. As an accessible alternative, ChatGPT, developed by OpenAI, has demonstrated the potential to assist in disease diagnosis through conversation with patients [22], [23], [24], [24], [25], [26]. By leveraging its advanced natural language processing capabilities, ChatGPT can interpret symptoms and medical history provided by patients and make suggestions for potential diagnoses or referrals to appropriate medical specialists. However, ChatGPT currently only allows text input and does not support direct image input for diagnosis, which limits its effectiveness in diagnosing medical conditions.

The idea of providing medical images directly for automatic diagnosis and generating text reports is exciting. ChatCAD [27] is one of the most advanced approaches that designed various networks to take X-rays, CT scans and MRIs images to generate diverse outputs, which are then transformed into text descriptions. These descriptions are combined as inputs to ChatGPT to generate a condensed report and offer interactive explanations and medical recommendations based on the given image. However, their proposed vision-text models were limited to their certain tasks. Meanwhile, for ChatCAD, users need to use ChatGPT's API to upload text descriptions, which could raise data privacy issues. To address those issues, MiniGPT-4 (https://minigpt-4.github.io) [28] is the first open-source method that allows users to deploy locally to interface images with LLMs and interact using natural language. However, MiniGPT-4 is trained on the combined dataset of Conceptual Caption [29], SBU [30] and LAION [31], which is irrelevant to medical images and, therefore, cannot be used for formal medical image diagnosis.

Here, we propose SkinGPT, the first dermatology diagnostic system that utilizes an advanced vision-based large language model. SkinGPT is the first system of its kind, incorporating a finetuned version of MiniGPT-4 with a vast collection of in-house skin disease images, accompanied by doctor's notes. With SkinGPT, users can upload their own skin photos for diagnosis, and the system can autonomously determine the characteristics and categories of skin conditions, perform analysis, and provide treatment recommendations. Unlike other diagnostic systems that require uploading images to a server or relying on ChatGPT, SkinGPT is designed to be deployed locally, thus ensuring the user's privacy. One of the key advantages of SkinGPT is that it empowers patients to gain a clearer understanding of their symptoms, diagnosis, and treatment plans. This improved understanding can help users engage in more effective and economical consultations with medical professionals. With SkinGPT, patients can have more informed conversations with their doctors, leading to better treatment outcomes

Fig. 2. Illustration of Fine-tuning Datasets. The notes below each image indicate clinical concepts as defined in **Method**.

and a higher level of satisfaction. In conclusion, SkinGPT represents a significant advancement in dermatology diagnostic systems. Its advanced vision-based large language model, coupled with a vast collection of in-house skin disease images, makes it a powerful tool for skin disease diagnosis and treatment. The ability to deploy it locally and protect user privacy makes SkinGPT an attractive option for patients seeking an accurate and reliable diagnosis of their skin conditions.

2 Results

2.1 The Overall Design of SkinGPT

SkinGPT is an interactive software designed to provide a text-based diagnosis of skin disease images as shown in Figure 1. The process begins with the user uploading a skin image, which is then processed by the Vision Transformer (VIT) and Q-Transformer models to understand its contents. The VIT model divides the image into smaller patches and extracts features such as edges, textures, and shapes. After that, the Q-Transformer model generates an embedding of the image based on the features identified by the VIT model. This is done using a transformer-based architecture that allows the model to consider the context of the image and generate an accurate diagnosis. The alignment layer then performs visual-text alignment, and the Vicuna outputs the text diagnosis. SkinGPT is fine-tuned on MiniGPT-4 using large in-house skin disease images along with doctor's notes to allow for medical diagnosis and interactive dialogue. The software can provide an interactive and user-friendly way to help users self-diagnose skin diseases, making it a useful tool for healthcare professionals and patients alike.

2.2 Interactive and Understandable Dermatology Diagnosis

SkinGPT offers numerous benefits to patients and medical professionals alike. One of its key advantages is its ability to leverage extensive and reliable medical knowledge to provide interactive explanations and recommendations. As shown in Figure 3, we can see two examples of interactive dialogue that demonstrate the capabilities of SkinGPT in action. In the top row of examples, an image is presented of an elderly man with actinic keratosis on his face. In the bottom row of examples, an image is provided of a patient with psoriasis on his hand.

With interactive dialogue, SkinGPT empowers patients with a better comprehension of their symptoms, diagnosis, and treatment plans, resulting in more productive and cost-effective consultations with medical professionals.

When we compare MiniGPT-4 and SkinGPT in Figure 3, it was observed that MiniGPT-4 did not provide accurate medical diagnosis, which can be attributed to its lack of optimization on medical datasets. However, SkinGPT was able to accurately identify the features of the skin diseases

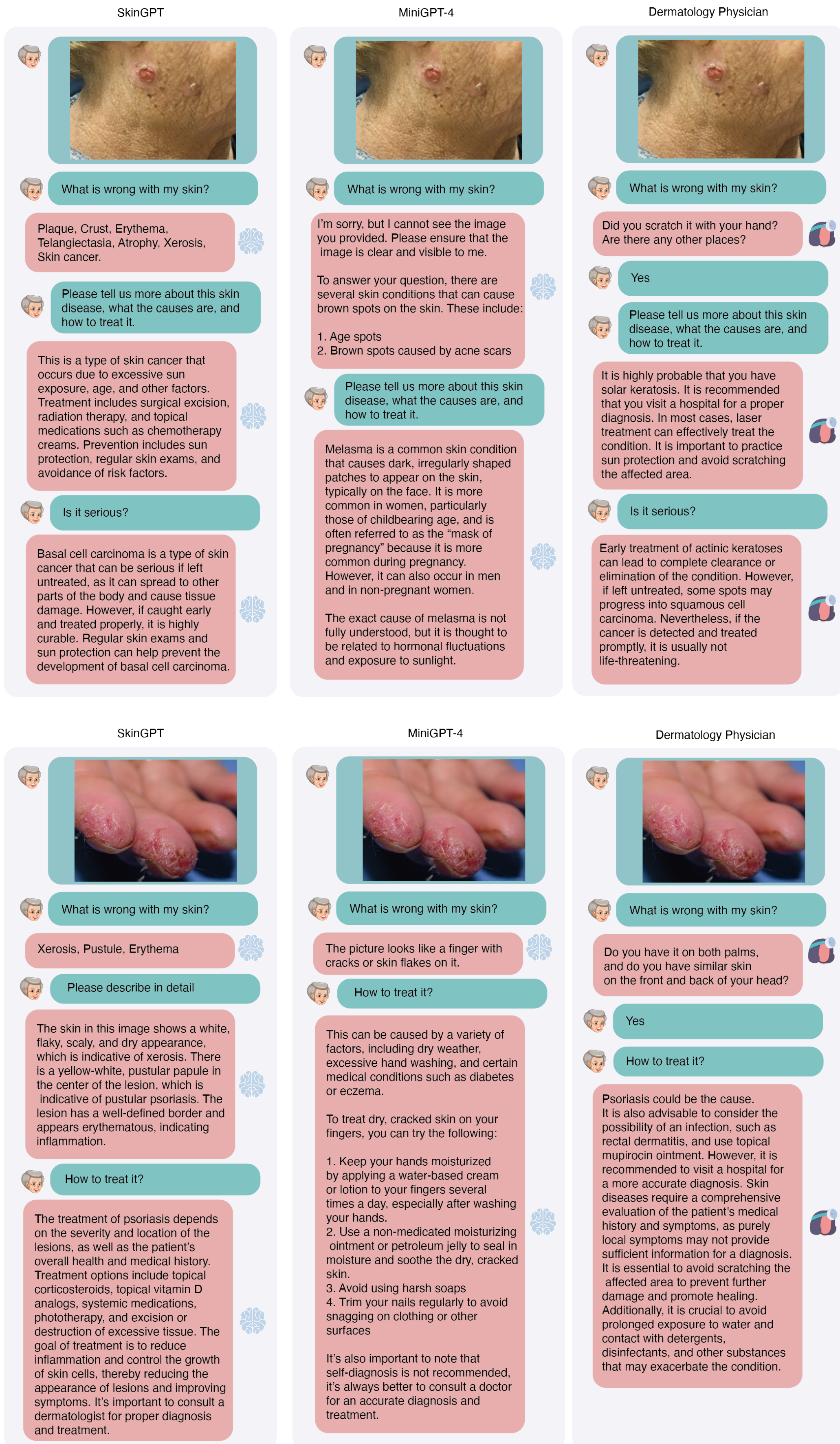

Fig. 3. Comparison between SkinGPT, MiniGPT-4 and Dermatology Physician.

depicted in the images, suggest possible disease types and propose potential treatments.

2.3 Physician's Evaluation of SkinGPT

In order to evaluate the reliability of SkinGPT, we conducted a study in which we compared its performance to that of online consultations with dermatologists across a large number of cases. The results of the study were compelling, as they showed that SkinGPT was able to provide an accurate diagnosis that was consistent with that of a dermatologist as shown in Figure 3.

In addition to being faster than online consultations as shown in Figure 4, SkinGPT also has the advantage of being available 24/7, making it a convenient option for patients who need a diagnosis outside of regular office hours. Moreover, SkinGPT's ability to provide a preliminary diagnosis can help patients make informed decisions about when and if they need to seek in-person medical attention. This can help reduce unnecessary trips to the doctor's office, saving patients time and money.

SkinGPT's potential to improve access to healthcare is especially important in rural areas or areas with a shortage of dermatologists. Patients in these areas may face long wait times or travel long distances to see a dermatologist. With SkinGPT, they can receive a preliminary diagnosis quickly and conveniently, potentially reducing the need for in-person visits. This can also help to alleviate the burden on healthcare systems in these areas.

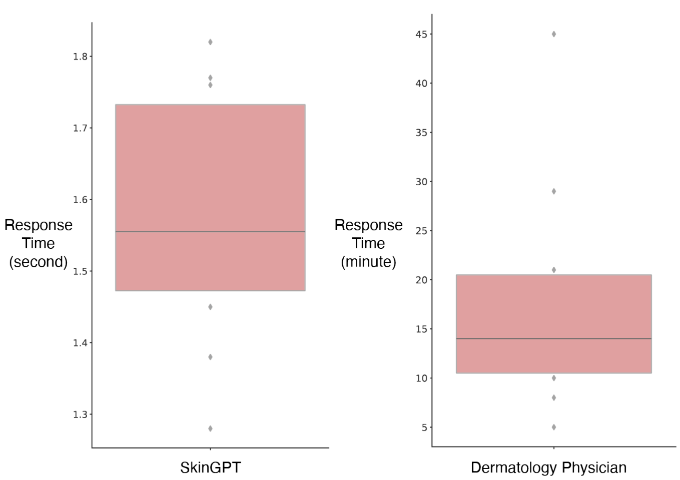

Fig. 4. Response time of SkinGPT compared to consulting dermatology physician online.

3 METHODS

3.1 Dataset

Our datasets include two public datasets and our private in-house dataset. The public dataset named SKINCON [32] is the first medical dataset densely annotated by domain experts to provide annotations useful across multiple disease processes. SKINCON is a skin disease dataset densely annotated by dermatologists and it includes 3230 images from the Fitzpatrick 17k skin disease dataset densely annotated with 48 clinical concepts, 22 of which have at least 50 images representing the concept, and 656 skin disease images from the Diverse Dermatology Images dataset. Our private in-house dataset contains 5620 images for 48 clinical concepts proposed in SKINCON by collecting either online or from hospitals. The 48 clinical concepts include Vesicle, Papule, Macule, Plaque, Abscess, Pustule, Bulla, Patch, Nodule, Ulcer, Crust, Erosion, Excoriation, Atrophy, Exudate, Purpura/Petechiae, Fissure, Induration, Xerosis, Telangiectasia, Scale, Scar, Friable, Sclerosis, Pedunculated, Exophytic/Fungating, Warty/Papillomatous, Dome-shaped, Flat-topped, Brown (Hyperpigmentation), Translucent, White (Hypopigmentation), Purple, Yellow, Black, Erythema, Comedo, Lichenification, Blue, Umbilicated, Poikiloderma, Salmon, Wheal, Acuminate, Burrow, Gray, Pigmented, and Cyst.

3.2 Finetuning of MiniGPT-4

SkinGPT is fine-tuned on MiniGPT-4 using large in-house skin disease images along with doctor's notes to allow for medical diagnosis and interactive dialogue. MiniGPT-4 aims to combine the power of a large language model with visual information obtained from a pre-trained vision encoder. To achieve this, the model uses Vicuna [33] as its language decoder, which is built on top of LLaMA [34] and is capable of performing complex linguistic tasks. To process visual information, the same visual encoder used in BLIP-2 [35] is employed, which consists of a ViT [36] backbone combined with a pre-trained Q-Former. Both the language and vision models are open-source.

To bridge the gap between the visual encoder and the language model, MiniGPT-4 utilizes a linear projection layer. An overview of the model is presented in Figure 1. Following the two-stage training approach proposed by MiniGPT-4. We fine-tune the pre-trained model provided by MiniGPT-4 with an image(skin disease images)-text(doctor's notes) dataset in the first stage. This helps to improve the model's generation reliability and usability for medical diagnosis. In the second stage, we fine-tuned the model using self-constructed datasets with medical images and refined corresponding doctor's notes to further enhance the model's performance in providing user-friendly reports.

3.3 Model Training and Resources

During model finetuning, the number of epochs was fixed at 20,000. The entire fine-tuning process takes about 1 hour to complete, utilizing 2 V100 (32GB) GPUs. The inference requires 2 V100(32GB) GPUs. The SkinGPT method was developed based on Python3.7, PyTorch1.9.1 and CUDA11.4. A detailed list of dependencies could be found in our code availability. A workstation with 252 GB RAM, 112 CPU cores and 2 Nvidia V100 GPUs was used.

4 Limitations and Discussion

Our study demonstrates that LLMs that utilize vision-based inputs have the potential to enhance medical diagnosis. With the upcoming release of more advanced LLMs like GPT-4, the accuracy and quality of diagnoses could be further improved. However, there are potential privacy concerns associated with using ChatGPT and GPT-4 as an API since users must upload data. Meanwhile, since users could deploy SkinGPT on their own, the privacy issue is solved.

Still, our work focused on qualitative analysis. Though we conducted comparisons and asked dermatologist for evaluations, we recognize that we cannot account for all cases of skin diseases worldwide. Therefore, we intend to address potential issues of SkinGPT while applying it in the real world and continue to gather more medical data to develop a more robust version.

As LLMs like SkinGPT continue to evolve and improve with the acquisition of even more reliable medical training data, the potential for significant advancements in online medical services is enormous. SkinGPT could play a critical role in improving access to healthcare and enhancing the quality of medical services for patients worldwide. We will continue our research in this field to further develop and refine this technology.

5 Acknowledgements

Funding: Juexiao Zhou and Xin Gao were supported in part by grants from the Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award number FCC/1/1976-44-01, FCC/1/1976-45-01, REI/1/5202-01-01, REI/1/5234-01-01, REI/1/4940-01-01, RGC/3/4816-01-01, and REI/1/0018-01-01.

Competing Interests: The authors have declared no competing interests.

Data availability: The de-identified teledermatology data used in this study are not publicly available due to restrictions in the data-sharing agreement.

Code availability: The SkinGPT is publicly available at https://github.com/JoshuaChou2018/SkinGPT.

References

[1] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf *et al.*, "The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions," *Journal of Investigative Dermatology*, vol. 134, no. 6, pp. 1527–1534, 2014.

[2] H. Feng, J. Berk-Krauss, P. W. Feng, and J. A. Stein, "Comparison of dermatologist density between urban and rural counties in the united states," *JAMA dermatology*, vol. 154, no. 11, pp. 1265–1271, 2018.

[3] J. Resneck Jr and A. B. Kimball, "The dermatology workforce shortage," *Journal of the American Academy of Dermatology*, vol. 50, no. 1, pp. 50–54, 2004.

[4] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature medicine*, vol. 26, no. 6, pp. 900–908, 2020.

[5] D. Seth, K. Cheldize, D. Brown, and E. E. Freeman, "Global burden of skin disease: inequities and innovations," *Current dermatology reports*, vol. 6, pp. 204–210, 2017.

[6] D. G. Federman, J. Concato, and R. S. Kirsner, "Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature," *Archives of family medicine*, vol. 8, no. 2, p. 170, 1999.

[7] G. Moreno, H. Tran, A. L. Chia, A. Lim, and S. Shumack, "Prospective study to assess general practitioners' dermatological diagnostic skills in a referral setting," *Australasian journal of dermatology*, vol. 48, no. 2, pp. 77–82, 2007.

[8] K. M. Yim, A. G. Florek, D. H. Oh, K. McKoy, and A. W. Armstrong, "Teledermatology in the united states: an update in a dynamic era," *Telemedicine and e-Health*, vol. 24, no. 9, pp. 691–697, 2018.

[9] P. R. Kshirsagar, H. Manoharan, S. Shitharth, A. M. Alshareef, N. Albishry, and P. K. Balachandran, "Deep learning approaches for prognosis of automated skin disease," *Life*, vol. 12, no. 3, p. 426, 2022.

[10] F. Martora, A. Ruggiero, G. Fabbrocini, and A. Villani, "Patient satisfaction with remote dermatology consultations during the covid-19 pandemic. comment on 'a qualitative assessment of patient satisfaction with remote dermatology consultations used during the uk's first wave of the covid-19 pandemic in a single, secondary-care dermatology department'," *Clinical and Experimental Dermatology*, vol. 47, no. 11, pp. 2037–2038, 2022.

[11] R. López-Liria, M. Á. Valverde-Martínez, A. López-Villegas, R. J. Bautista-Mesa, F. A. Vega-Ramírez, S. Peiró, and C. Leal-Costa, "Teledermatology versus face-to-face dermatology: An analysis of cost-effectiveness from eight studies from europe and the united states," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, p. 2534, 2022.

[12] N. Lakdawala, C. Gronbeck, and H. Feng, "Workforce characteristics of nonphysician clinicians in dermatology in the united states," *Journal of the American Academy of Dermatology*, vol. 87, no. 5, pp. 1108–1110, 2022.

[13] I. K. Pious and R. Srinivasan, "A review on early diagnosis of skin cancer detection using deep learning techniques," in *2022 International Conference on Computer, Power and Communications (ICCPC)*. IEEE, 2022, pp. 247–253.

[14] A. A. Cruz-Roa, J. E. Arevalo Ovalle, A. Madabhushi, and F. A. González Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer, 2013, pp. 403–410.

[15] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[16] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The lancet oncology*, vol. 20, no. 7, pp. 938–947, 2019.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[18] S. S. Han, G. H. Park, W. Lim, M. S. Kim, J. I. Na, I. Park, and S. E. Chang, "Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network," *PloS one*, vol. 13, no. 1, p. e0191493, 2018.

[19] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 206–222.

[20] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, p. e0000198, 2023.

[21] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[22] M. Balas and E. B. Ing, "Conversational ai models for ophthalmic diagnosis: Comparison of chatgpt and the isabel pro differential diagnosis generator," *JFO Open Ophthalmology*, vol. 1, p. 100005, 2023.

[23] M. Mijwil, M. Aljanabi, and A. H. Ali, "Chatgpt: Exploring the role of cybersecurity in the protection of medical information," *Mesopotamian journal of cybersecurity*, vol. 2023, pp. 18–21, 2023.

[24] R. K. Sinha, A. D. Roy, N. Kumar, H. Mondal, and R. Sinha, "Applicability of chatgpt in assisting to solve higher order problems in pathology," *Cureus*, vol. 15, no. 2, 2023.

[25] F. Ufuk, "The role and limitations of large language models such as chatgpt in clinical settings and medical journalism," *Radiology*, p. 230276, 2023.

[26] M. Hu, S. Pan, Y. Li, and X. Yang, "Advancing medical imaging with language models: A journey from n-grams to chatgpt," *arXiv preprint arXiv:2304.04920*, 2023.

[27] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[28] D. Zhu, J. Chen, X. Shen, xiang Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.

[29] P. Sharma, N. Ding, S. Goodman, , and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, 2018*.

[30] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[31] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[32] R. Daneshjou, M. Yuksekgonul, Z. R. Cai, R. Novoa, and J. Y. Zou, "Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 157–18 167, 2022.

[33] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, , and E. P. Xing., "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023."

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[35] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[36] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," *arXiv preprint arXiv:2211.07636*, 2022.