DermAI: A Chatbot Assistant for Skin lesion Diagnosis Using Vision and Large Language Models

Viet-Tham Huynh[1,2♣], Trong-Thuan Nguyen[1,2,3♣], Thao Thi-Phuong Dao[4], Tam V. Nguyen[5], and Minh-Triet Tran[1,2♠]

[1] Software Engineering Laboratory and Faculty of Information Technology University of Science, VNU-HCM, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Department of Electrical Engineering and Computer Science, University of Arkansas, U.S.A.
[4] Department of Otolaryngology, Thong Nhat Hospital, Ho Chi Minh City
[5] Department of Computer Science, University of Dayton, U.S.A.
{hvtham,ntthuan,tmtriet}@selab.hcmus.edu.vn, thao.dao2020@ict.jvn.edu.vn, tamnguyen@udayton.edu

Abstract. In dermatology, the demand for accurate skin lesion diagnoses is critical, especially during peak times like summer when skin cancer screenings surge. The need for efficient processing of large volumes of medical images and the risk of human error highlights the importance of innovative diagnostic tools. In this paper, we propose DermAI, an advanced AI-driven framework to improve diagnostic accuracy and efficiency in skin lesion analysis. Our DermAI framework combines a state-of-the-art segmentation model and a large language model to assist clinicians in interpreting medical images swiftly and precisely. Our framework isolates and analyzes key lesion features using advanced segmentation models and vision encoders, while a large language model provides contextual insights to understand lesion characteristics and potential malignancies. By integrating visual and linguistic analysis, our DermAI framework reduces diagnostic errors, alleviates clinician workloads, and enhances patient care with faster, more accurate results, supporting dermatologists in making informed decisions and advancing AI-assisted diagnostics.

Keywords: Large Language Model · Medical Vision-Language Model · Skin lesion Segmentation

1 Introduction

In today's healthcare landscape, dermatologists and clinicians face increasing pressure to interpret complex medical images, such as skin lesion scans, within limited timeframes. During peak seasons, particularly in summer, when skin cancer screenings surge, clinicians are tasked with reviewing hundreds of lesion

♣ Equal Contribution ♠ Corresponding Author

images daily, where even minor delays in diagnosis can significantly impact patient outcomes [13, 22]. These high-stakes situations increase the risk of human error or oversight, placing additional pressure on healthcare professionals.

The growing demand for enhanced diagnostic accuracy and efficiency calls for advanced tools in medical imaging. Artificial Intelligence (AI) [23, 28] offers a transformative solution, reshaping how medical images are analyzed. To this end, we address this need through an AI-driven chatbot that integrates advanced vision and language models to assist in skin lesion diagnosis. The objective is to ease the workload on healthcare professionals, minimize diagnostic errors, and deliver faster, more accurate insights, ultimately improving patient care. In this paper, we present ***DermAI*** (Dermatology AI Assistant), a novel framework that utilizes a segmentation model to isolate key areas of skin lesions, focusing on clinically relevant features. The vision component, powered by a state-of-the-art medical image segmentation model [6] and vision encoder [26, 31], processes visual data, while a linear transformation enhances image quality for more accurate interpretation. In addition, based on Generative Pre-trained Transformer 4 (GPT-4) [17], the language component contextualizes the visual data, providing deeper insights into lesion characteristics and potential malignancy. Our approach delivers comprehensive analyses, supporting healthcare professionals in making well-informed decisions.

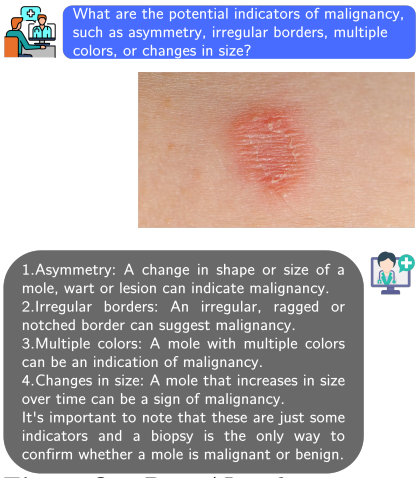

Fig. 1: Our DermAI with an interactive chat assistant. The user enters a prompt and a skin photo. The assistant responds with text feedback.

The Contributions of this Work. We first investigate state-of-the-art skin lesion segmentation methods. Then, we introduce a novel framework integrating vision with a Large Language Model (LLM) trained on large-scale medical data, namely ***DermAI***, illustrated in Fig. 1. Our experiments show exceptional performance in skin lesion segmentation and practical interpretation, earning recognition from medical professionals.

2 Related Work

2.1 Skin Lesion Segmentation

Recently, medical image segmentation methods inspired by Transformer [16] and CNN architectures [20]. In particular for skin lesion segmentation methods (e.g., MobileUNetR [19] and DuAT [25]), which balance efficiency and accuracy by addressing the challenges of preserving global context and local detail. Polar transformations [5] and boundary-aware mechanisms [30] also enhance segmentation performance and data efficiency by capturing crucial boundary information.

In addition, the integrating diffusion models (e.g., DermoSegDiff [6] and MedSegDiff [34]) with attention mechanisms further improve boundary delineation, which improving overall segmentation quality. To tackle data imbalance, particularly in small lesion segmentation, the Focal Tversky loss function [1] improves the precision-recall trade-off. At the same time, Inconsistency Masks (IM) [29] enable strong results with minimal labeled data. Moreover, enhanced U-Net variants such as DoubleU-Net [12] and BCDU-Net [4], which utilize pre-trained encoders, dense connections, and bi-directional ConvLSTM for feature extraction and fusion. Multi-scale approaches like MSRF-Net [24] and context-gating mechanisms [3] address variable object sizes and complex anatomical variations.

2.2 Medical Chatbot

The introduction of large language models, particularly ChatGPT, has ignited increasing interest in developing medical chatbots, especially for their capacity to automate X-ray image analysis. These technologies serve as valuable tools for patients and healthcare professionals by facilitating a more comprehensive understanding of diagnostic findings from X-ray images. Recent advancements in Large Language Models (LLMs) and Multi-Modal Learning have highlighted the potential of these systems in medical applications. Several notable works have emerged in this domain, including Chatdoctor [15], LLaMA [27], MedAlpaca [11], PMC-LLaMA [33], and DoctorGLM [35]. For instance, Chatdoctor [15], built upon the LLaMA [27] model, provides reliable interpretations of X-ray images for both patients and clinicians, offering personalized medical advice. Similarly, MedAlpaca [11], PMC-LLaMA [33], and DoctorGLM [35] have fine-tuned open-source LLMs on medical data to develop chatbots tailored to healthcare contexts. These advancements emphasize the growing potential of integrating LLMs and multi-modal learning into medical applications, paving the way for more personalized, accurate, and accessible diagnostic tools in healthcare.

2.3 Discussion

As presented in Sections 2.1 and 2.2, while advancements in skin lesion segmentation and medical chatbots have progressed, a noticeable gap remains in integrating chatbots with skin lesion segmentation capabilities. Although chatbots have proven valuable in tasks like X-ray analysis, none currently harness the power of segmentation for skin lesions, a crucial tool in dermatology. Developing a DermAI chatbot, which combines skin lesion segmentation with interactive conversational capabilities, offers transformative potential. Our system assists healthcare professionals in making precise diagnoses by delivering real-time segmented imagery with expert guidance while empowering patients with personalized, easily understandable feedback. This innovation has the potential to streamline dermatological workflows, boost diagnostic accuracy, improve patient outcomes, and bridge a gap in healthcare technology for dermatology.

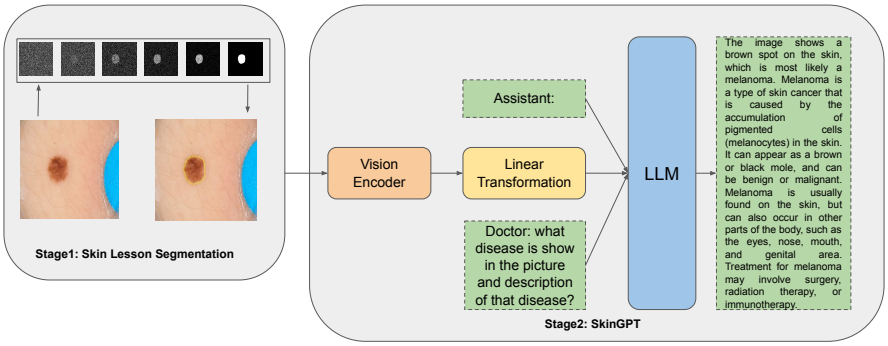

Fig. 2: Overview of the proposed framework, namely DermAI. Our approach consists of two stages: skin lesion segmentation and SkinGPT. In the first stage, the segmentation model [6] isolates the lesion from the input image. In the second stage, SkinGPT [37] leverages a pre-trained Vision Transformer and LLM (Llama-2 [27]) to provide context-aware diagnostic insights based on the segmented image.

3 Our Proposed Framework

Framework Overview. Our proposed framework, illustrated in Fig. 2, introduces a two-stage process designed to revolutionize dermatological diagnostics. In the first stage, we leverage DermoDiff [6] to segment the input image, isolating key features for analysis. This segmented output is fed into SkinGPT [37], an innovative, multimodal diagnostic system powered by large language models. By aligning a pre-trained Vision Transformer with the Llama-2 LLM, SkinGPT leverages an extensive dataset of skin disease images enriched with clinical concepts and doctors' notes to generate highly insightful diagnostic outputs.

3.1 Skin lesion Segmentation

Skin lesion segmentation plays a critical role in medical imaging and dermatology, serving as a cornerstone in diagnosing and analyzing various skin conditions, including life-threatening cancers such as melanoma. With the global incidence of skin cancer rising and melanoma being one of the most aggressive forms, the demand for accurate, automated diagnostic tools has grown significantly. Segmentation techniques are pivotal in this context, enabling healthcare professionals to identify, assess, and monitor skin lesions from dermoscopic or clinical images, supporting more timely and informed medical decisions. In our approach, we levage DermoSegDif [6] including ***an encoder***, ***a bottleneck***, and ***a decoder***.
Encoder. The Encoder consists of a series of stacked Encoder Modules (EM), followed by a convolution layer that reduces the spatial dimensions to a four-by-four tensor. Instead of the conventional approach of concatenating ϵ_θ and g_{i-1} before feeding them into the network, as proposed in prior work [32], the authors introduce a two-path feature extraction strategy within each EM. This method emphasizes the mutual influence between the noisy segmentation mask and the

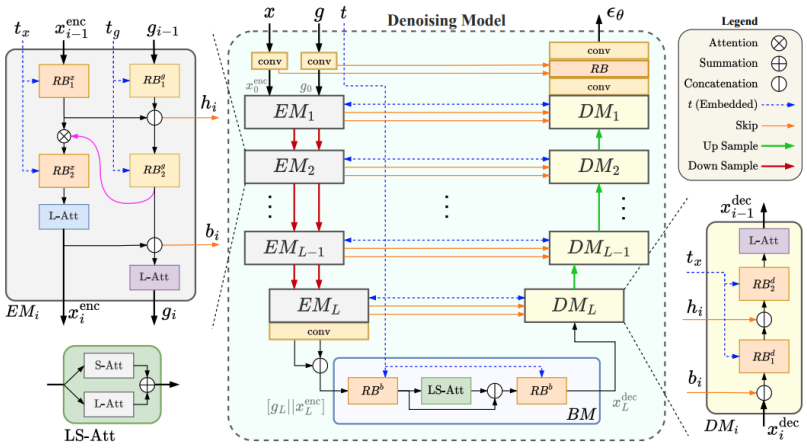

Fig. 3: Overview of the DermoSegDif [6] method.

guidance image. Each path in the encoder includes two ResNet Blocks (RB) and a Linear Attention (L-Att) mechanism, providing computational efficiency and non-redundant feature extraction. Time embeddings are incorporated into each RB using sinusoidal positional embeddings, processed through linear layers and GeLU activation functions. Separate time embeddings are used for the guidance image (t_g) and the noisy segmentation mask (t_x), allowing the model to capture the temporal dynamics of both inputs. To enhance feature extraction, knowledge from the noise path (RB_1^x) is transferred and concatenated with the guidance path, creating an intermediate feature (h_i) that captures complementary representations. The guidance path processes the output through RB_2^g, and a feedback mechanism applies a convolution to the output, reconnecting it to RB_2^x. This feedback loop ensures boundary and noise information integration, allowing the model to emphasize key features while suppressing irrelevant details.

Bottleneck. The final outputs of the encoder, x_L^{enc} and g_L, are concatenated and passed through the Bottleneck Module (BM). This module includes a ResNet Block (RB), a Linear Self-Attention (LS-Att) mechanism, and another ResNet Block. The LS-Att module enhances feature representation by combining the spatial relationships captured by Self-Attention (S-Att) and the semantic context captured by Linear Attention (L-Att). These two attention mechanisms operate in parallel, allowing the model to integrate spatial and contextual information effectively. The output from the Bottleneck Module is then passed to the decoder.

Decoder. The Decoder comprises stacked Decoder Modules (DM) that match the number of Encoder Modules (EM). Each DM operates as a single-path module, consisting of two consecutive ResNet Blocks (RB) and one Linear Attention (L-Att) module, followed by a convolutional block that outputs the estimated noise ϵ_θ. The decoder integrates information from both the noise and guidance paths by concatenating the encoder outputs, b_i and h_i, before and after applying RB_1^d. This enables the decoder to effectively utilize the refined features from the encoder, improving its ability to estimate the added noise and recover missing

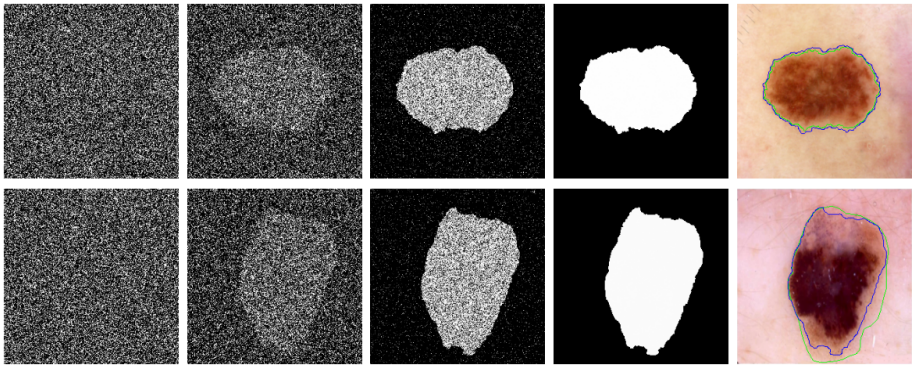

Fig. 4: Illustration of the diffusion process: noise is progressively added to the input image and then estimated and reduced by the model to reconstruct an accurate segmentation mask, delineating the boundaries of the lesion.

segmentation details. A skip connection is also introduced, linking the original input x to the final decoder layer. This skip connection concatenates the output of the first decoder module (DM_1) with x_1, and the combined features are processed through a final convolutional block to produce the estimated noise ϵ_θ.

Fig. 4 illustrates the diffusion process, a key component in generating and refining noise for reconstructing segmentation masks. In this process, noise is incrementally added to the input image to simulate data uncertainty, which the model progressively estimates and reduces to recover the lesion's clear boundaries. The model refines its predictions as noise is removed, producing an accurate segmentation mask that outlines the lesion's exact contours. A skip connection linking the original input to the final decoder layer preserves critical details, combining them with intermediate features to enhance segmentation. The model effectively handles complex, irregular lesion boundaries by predicting the noise (ϵ_θ) at each step, resulting in more precise and robust segmentation.

3.2 SkinGPT

SkinGPT, built on ChatGPT [17] and fine-tuned for dermatology, has gained attention for its ability to assist healthcare professionals in analyzing and interpreting medical images. A key application involves integrating a skin lesion segmentation model with ChatGPT to enhance diagnostic accuracy when processing medical images and responding to prompts, such as doctors' inquiries.

In the first stage, MiniGPT-4 [38] is trained to understand the alignment between visual information and language by learning from a large dataset of image-text pairs. It uses Vicuna [18], a language decoder based on LLaMA [27], and adopts the Vision Transformer (ViT) [2] from BLIP-2 [14], extracting features with Q-Former [36]. In addition, a linear projection layer is introduced to align the features from the visual encoder with the LLM, transforming them into soft prompts for generating textual descriptions. During pretraining, only the projection layer is trained while the vision encoder and LLM remain frozen.

In the second stage, the model undergoes fine-tuning to address issues from the first stage, such as incoherent outputs and repetitive phrases. Since high-quality vision-language datasets are scarce, the authors created their own by generating detailed image descriptions using the pre-trained model. Prompts were designed to encourage exhaustive descriptions in a conversational format, such as:

```
###Human:<Img><ImgFeature></Img> Describe this image in detail.
###Assistant:
```

If descriptions were too short, additional prompts were used to elicit more comprehensive responses. Despite generating a large number of image-text pairs, many descriptions still contained errors like redundancy or irrelevant content. ChatGPT was used to clean the data, automate refinement, and remove errors and redundant information. Moreover, MiniGPT-4 is fine-tuned using the curated dataset. The predefined prompt template follows the below format:

```
###Human: <Img><ImageFeature></Img <Instruction>
###Assistant:
```

where `<Instruction>` refers to a randomly selected prompt, such as "*Describe this image in detail*" or "*Could you describe the contents of this image for me?*". The fine-tuning process aims to enhance the model's ability to generate natural, coherent language that aligns contextually with the visual input. Importantly, no regression loss is calculated for these text-image prompts, as the primary objective is to improve the fluency and reliability of the model's output.

3.3 DermAI: Dermatology AI Assistant

Our DermAI system first processes the input image by isolating the region of interest (ROI) through a segmentation method that creates a mask M, highlighting the skin lesions. This mask is then applied to the image, producing a masked version I_M that reveals only the lesion area. Next, the masked image I_M is analyzed by SkinGPT, a language model fine-tuned for dermatology, which examines key features such as the lesion's shape, color, borders, and texture. By focusing solely on the segmented region, DermAI ensures attention to clinically significant areas, generating detailed, context-aware explanations.

Discussion. DermAI's integration of segmentation delivers significant advantages by focusing the model's attention on the region of interest (ROI), allowing for a more precise and detailed analysis of the lesion. This targeted approach enhances the ability of SkinGPT to identify critical lesion characteristics, such as asymmetry, border irregularity, and color variation, which are essential for diagnosing conditions like melanoma. Our framework minimizes distractions by filtering out irrelevant areas, leading to more specific and clinically relevant insights, ultimately improving the reliability of DermAI's diagnostic outputs.

Segmentation and SkinGPT work together within our DermAI framework to maximize precision and contextual relevance in dermatological analysis. Segmentation ensures the model concentrates on the most important regions of the image, while SkinGPT, fine-tuned for medical language tasks, generates detailed, context-aware explanations. Our approach enables more accurate diagnoses by

leveraging the strengths of both components, ultimately providing healthcare professionals with reliable and actionable information for evaluating skin lesions.

4 Experiment Results

4.1 Implementation Details

Training. We first train a segmentation model on the ISIC [8, 10] datasets. The segmented images are then analyzed with the pre-trained SkinGPT [10], which was initially trained on SKINCON [9], a dataset densely annotated by dermatologists across multiple skin disease concepts, and Dermnet, which includes a diverse range of images features skin disease classes by board-certified dermatologists.
Inference. During inference, DermAI operates as a unified framework by processing the skin image and the user-provided prompt. Our system uses the segmentation model to isolate the skin lesion, generating a masked image. This segmented image and the prompt are fed into the SkinGPT model, where the visual data and the prompt are jointly analyzed to produce a response.
Datasets. The *ISIC 2016* [10] dataset includes 900 training images and 379 test images, each with expert-annotated binary masks delineating lesion boundaries. The *ISIC 2017* [8] dataset includes 2,000 training images and 150 validation images. Additionally, the SkinGPT model is trained on two large-scale datasets. While the SKINCON [9] dataset features diverse dermatological annotations across 48 clinical concepts, the Dermnet[6] dataset covers 15 skin diseases.
Evaluation Metrics. We evaluate the segmentation model using the *Dice* and Intersection over Union (IoU) as previous work for fair comparison. Higher Dice values reflect better region overlap, indicating improved segmentation accuracy. Higher IoU values denote closer alignment between predictions and ground truth.

4.2 Quantitative Analysis

The results in Table 1 compare Dice scores across the ISIC 2016 and ISIC 2017 datasets, highlighting the performance variation of DermoDiff against baseline methods. On the ISIC 2016 dataset, DermoDiff achieves a competitive Dice score of 90.37%, closely following the highest scorer, UNet, which reaches 89.84%. Our experimental results suggest that DermoDiff is well-suited for handling relatively straightforward lesion segmentation tasks, where the boundaries between lesions and healthy tissue are well-defined, making segmentation less complex.

Table 1: Comparison (%) on ISIC 2016 and ISIC 2017 against baseline methods at *Dice*.

Methods	ISIC'16	ISIC'17
Swin-UNet [7]	85.68	**79.14**
UNet [21]	89.84	77.08
DermoDiff [6]	**90.37**	74.63

[6] https://www.kaggle.com/datasets/shubhamgoel27/dermnet

However, the performance gap becomes more pronounced when analyzing the ISIC 2017 dataset. The Dicde score of DermoDiff drops to 74.63%, notably behind Swin-UNet at 79.14%. This decrease indicates that DermoDiff struggles with the more challenging images in the ISIC 2017 dataset, where lesions are often more irregular in shape, size, and texture and may present higher variability in appearance. For example, cases require advanced feature extraction and spatial context understanding, which DermoDiff appears to have limitations in addressing effectively. The lower score of DermoDiff points to a need for enhancing the ability to generalize across diverse datasets, particularly when dealing with more nuanced clinical cases where lesion characteristics are less homogeneous.

Table 2 further reinforces these observations through IoU scores, which offer an additional perspective on model accuracy in identifying lesion boundaries. DermoDiff performs admirably on the ISIC 2016 dataset, securing an IoU of 82.43%, slightly outperforming UNet (82.09%) and significantly ahead of Swin-UNet (75.59%). This strong performance on ISIC 2016 underscores the capability of DermoDiff in environments where segmentation challenges are moderate and lesions are relatively distinguishable. However, the drop in IoU to 59.53% on the ISIC 2017 dataset is significant, highlighting the limitations of DermoDiff in more intricate segmentation tasks. This decline suggests that DermoDiff still struggles with distinguishing between lesion boundaries and surrounding tissues when faced with more significant variability in lesion presentations.

Table 2: Comparison (%) on ISIC 2016 and ISIC 2017 against baseline methods at *IoU*.

Methods	ISIC'16	ISIC'17
Swin-UNet [7]	75.59	**66.76**
UNet [21]	82.09	64.10
DermoDiff [6]	**82.43**	59.53

Additionally, the disparity in results between the ISIC 2016 and ISIC 2017 datasets suggests that DermoDiff is optimized for cases where lesion appearances are more consistent and well-defined, such as those on ISIC 2016, but performs less effectively with the more heterogeneous cases on the ISIC 2017. This underperformance on ISIC 2017 may stem from its reduced ability to capture complex spatial relationships or its reliance on prominent features in simpler images. In contrast, the superior performance of Swin-UNet on ISIC 2017 highlights that models using advanced transformer-based architectures are better equipped to handle complex skin lesion segmentation, as they can capture long-range dependencies and contextual information more effectively.

4.3 Qualitative Analysis

We provide qualitative evaluations of the segmentation model in Fig. 5, illustrating the precision of the predictions (blue outlines) compared to the ground truth (green outlines) across a spectrum of skin lesions. The model exhibits commendable accuracy in lesions with homogeneous boundaries, such as in the top left and bottom center images, where the blue and green outlines align closely. However, it encounters challenges with lesions that exhibit more complex features

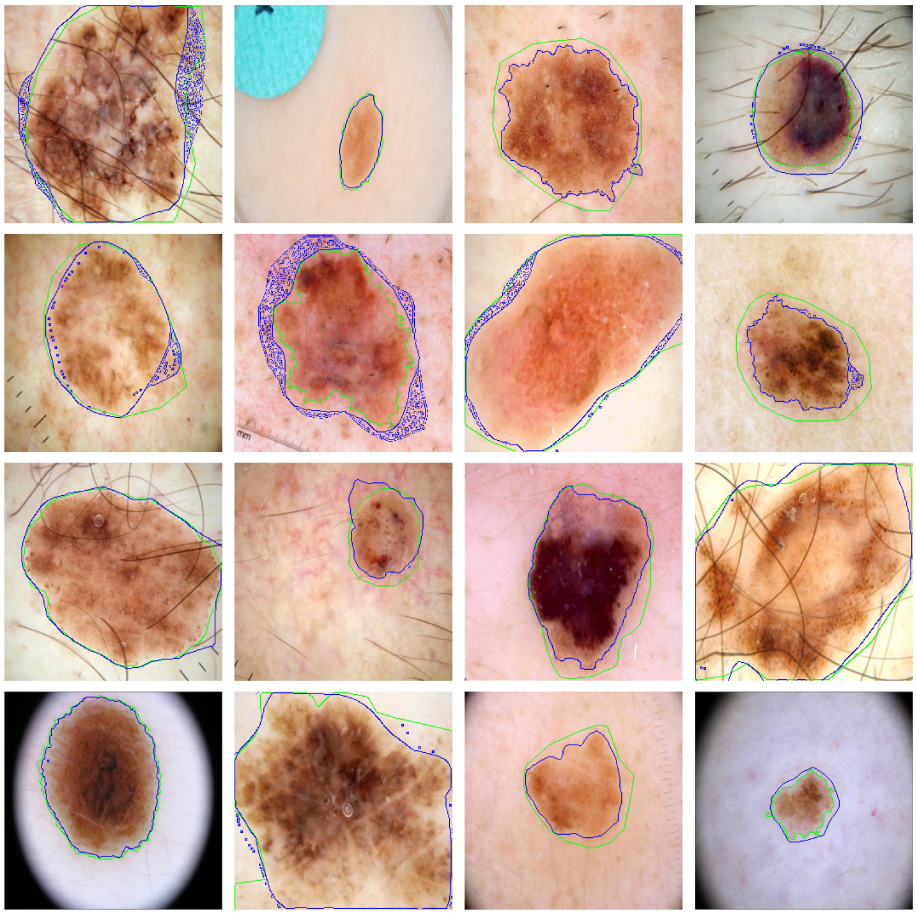

Fig. 5: Comparison of skin lesion segmentation results: Blue outlines show model predictions, while green outlines represent the ground truth.

like irregular borders or heterogeneous pigmentation. For instance, the image in the top second from left shows the model overextending the lesion boundary, a common issue in cases where the lesion fades gradually into surrounding tissue, leading to over-segmentation. In contrast, the bottom second from the left image highlights under-segmentation, where the model fails to capture the entire lesion, likely due to its inability to detect subtle differences in color and texture.

In addition, Fig. 6 underscores the critical importance of segmentation in medical image analysis, particularly when leveraged by the advanced capabilities of DermAI. Moreover, Fig. 1 illustrates the interaction of the DermAI chat assistant with an internet-sourced image (not from the ISIC 2016 or 2017 datasets), demonstrating the generalization of our DermAI framework. By isolating regions of interest, DermAI can focus its pattern recognition algorithms on specific areas, identifying subtle dermatological markers such as asymmetry,

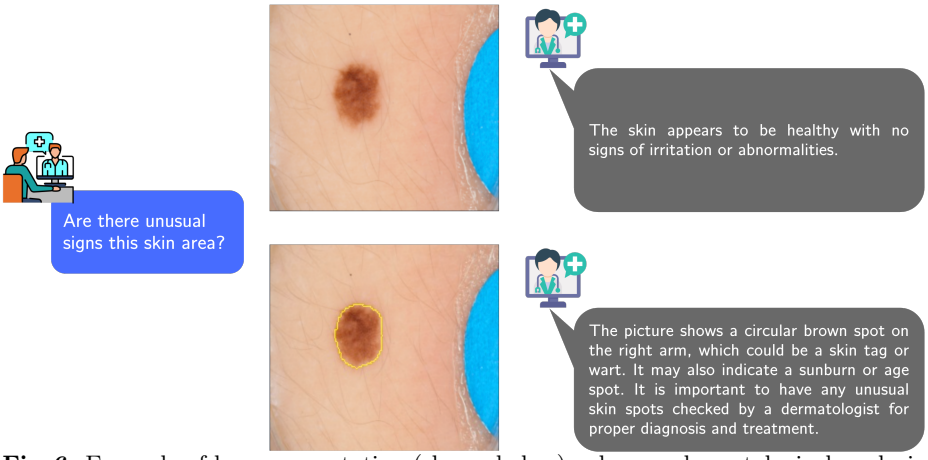

Fig. 6: Example of how segmentation (shown below) enhances dermatological analysis by highlighting a skin lesion for focused and precise assessment by DermAI.

irregular borders, color variation, and changes in size, which are factors crucial for distinguishing malignant lesions, like melanoma, from benign growths such as warts and skin tags. This segmentation-driven approach closely mimics the systematic evaluation process employed by dermatologists, ensuring that the model's analysis adheres to clinical best practices. DermAI effectively reduces image noise, enhances diagnostic accuracy, and delivers clinically relevant insights by isolating and analyzing distinct, well-defined regions. The refined focus of our proposed approach significantly enhances the precision of lesion identification by concentrating on the most relevant areas of interest. Therefore, DermAI leads to more accurate and reliable diagnoses, enabling healthcare professionals to make informed, actionable decisions. By reducing false positives and increasing diagnostic clarity, our approach contributes to higher-quality assessments in dermatological care, ultimately supporting improved patient outcomes.

Expert Medical Evaluation. In a comprehensive evaluation, we presented our software to three medical experts, who compared two versions of DermAI: one with segmentation and one without. All three unanimously agreed that the version with segmentation significantly outperformed the one without. The segmentation-enhanced model consistently identified critical lesion features such as shape, color, borders, and location while providing more actionable insights regarding potential causes and recommendations for diagnosis or treatment. This feedback demonstrates that segmentation enhances the model's ability to detect key visual features and improves the accuracy of its diagnostic insights.

In contrast, the version without segmentation struggled to produce focused and detailed responses, often lacking the specificity required for accurate diagnosis. Segmentation empowered the model to concentrate on the most relevant areas of the lesion, resulting in a more thorough and precise analysis. Experts noted that the segmented model delivered far more informative and clinically useful descriptions. Based on this feedback, we will include additional lesion at-

tributes such as size, surface texture, and fluid presence, which will further refine diagnostic accuracy. Importantly, these evaluations underscore the importance of segmentation in boosting diagnostic precision and clinical effectiveness.

5 Conclusion

In this paper, we have introduced DermAI, a novel framework designed to enhance the efficiency of skin lesion diagnosis by integrating advanced vision and language models. Our system combines state-of-the-art segmentation models with LLM, allowing clinicians to swiftly and accurately interpret medical images while reducing the risk of diagnostic errors. Integrating vision encoders with language models provides a comprehensive understanding of skin lesions, helping dermatologists make well-informed decisions. The results demonstrate the effectiveness of DermAI in streamlining diagnostic workflows and improving patient care, particularly during peak times, such as in summer when skin cancer screenings surge. Despite its promising capabilities, DermAI also highlights the importance of continuous refinement to better generalize across diverse datasets, as shown by performance variations between ISIC 2016 and ISIC 2017 datasets.

In the future, we aim to enhance the robustness and clinical utility of DermAI. First, we will refine the segmentation model to handle complex and diverse datasets, improving its generalization across different lesion types and skin conditions. Second, we will incorporate active learning strategies, enabling the model to continuously learn and improve from real-world data with minimal human intervention. Third, we plan to expand DermAI's capabilities to support real-time image processing and provide clinicians with interactive, interpretable AI-driven insights. Lastly, future iterations will explore integrating multimodal data inputs, such as patient history or genetic information, to create a more holistic diagnostic tool, further increasing precision and clinical relevance.

Acknowledgements. This research is funded by Viet Nam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01. We also acknowledge Dr. Cao Nu Hoang Oanh, Specialist Level I in Dermatology, for invaluable discussions and feedback.

References

1. Abraham, N., Khan, N.M.: A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 683–687. IEEE (2019) 3
2. Alexey, D.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020) 6
3. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020) 3
4. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019) 3
5. Benčević, M., Galić, I., Habijan, M., Babin, D.: Training on polar image transformations improves biomedical image segmentation. IEEE access **9**, 133365–133375 (2021) 2
6. Bozorgpour, A., Sadegheih, Y., Kazerouni, A., Azad, R., Merhof, D.: Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. In: International Workshop on PRedictive Intelligence In MEdicine. pp. 146–158. Springer (2023) 2, 3, 4, 5, 8, 9
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 8, 9
8. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018) 8
9. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.Y.: Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. Advances in Neural Information Processing Systems **35**, 18157–18167 (2022) 8
10. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 (2016) 8
11. Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023) 3
12. Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS). pp. 558–564. IEEE (2020) 3
13. Jones, O.T., Ranmuthu, C.K., Hall, P.N., Funston, G., Walter, F.M.: Recognising skin cancer in primary care. Advances in therapy **37**(1), 603–616 (2020) 2
14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 6

15. Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023) 3
16. Nguyen, T.T., Nguyen, T.V., Tran, M.T.: Collaborative consultation doctors model: Unifying cnn and vit for covid-19 diagnostic. IEEE Access (2023) 2
17. OpenAI: Gpt-4 technical report (2023) 2, 6
18. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023) 6
19. Perera, S., Erzurumlu, Y., Gulati, D., Yilmaz, A.: Mobileunetr: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. arXiv preprint arXiv:2409.03062 (2024) 2
20. Phung, K.A., Nguyen, T.T., Wangad, N., Baraheem, S., Vo, N.D., Nguyen, K.: Disease recognition in x-ray images with doctor consultation-inspired model. Journal of Imaging **8**(12), 323 (2022) 2
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015) 8, 9
22. Soenksen, L.R., Kassis, T., Conover, S.T., Marti-Fuster, B., Birkenfeld, J.S., Tucker-Schwartz, J., Naseem, A., Stavert, R.R., Kim, C.C., Senna, M.M., et al.: Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Science Translational Medicine **13**(581), eabb3652 (2021) 2
23. Srivastav, S., Chandrakar, R., Gupta, S., Babhulkar, V., Agrawal, S., Jaiswal, A., Prasad, R., Wanjari, M.B.: Chatgpt in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus **15**(7) (2023) 2
24. Srivastava, A., Jha, D., Chanda, S., Pal, U., Johansen, H.D., Johansen, D., Riegler, M.A., Ali, S., Halvorsen, P.: Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(5), 2252–2263 (2021) 3
25. Tang, F., Xu, Z., Huang, Q., Wang, J., Hou, X., Su, J., Liu, J.: Duat: Dual-aggregation transformer network for medical image segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 343–356. Springer (2023) 2
26. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023) 2
27. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 3, 4, 6
28. Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., et al.: Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654 (2024) 2
29. Vorndran, M.R., Roeck, B.F.: Inconsistency masks: Removing the uncertainty from input-pseudo-label pairs. arXiv preprint arXiv:2401.14387 (2024) 3
30. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 206–216. Springer (2021) 2
31. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022) 2

32. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: Konukoglu, E., Menze, B., Venkataraman, A., Baumgartner, C., Dou, Q., Albarqouni, S. (eds.) Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, pp. 1336–1348. PMLR (06–08 Jul 2022), https://proceedings.mlr.press/v172/wolleb22a.html 4
33. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454 (2023) 3
34. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024) 3
35. Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D.: Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097 (2023) 3
36. Zhang, Q., Zhang, J., Xu, Y., Tao, D.: Vision transformer with quadrangle attention. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 6
37. Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., et al.: Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. Nature Communications **15**(1), 5649 (2024) 4
38. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=1tZbq88f27 6