Original Research Article

AutoLesion: Accessible AI-Based Classification of Skin-Lesions Using Custom Vision Language Models

Shreya Talukder

Winston Churchill High School, 11300 Gainsborough Rd, Potomac, MD 20854, United States

ABSTRACT

Skin cancer is the most common form of cancer worldwide, and early detection plays a critical role in improving survival outcomes. While early-stage melanoma has a five-year survival rate of approximately 99%, this rate declines significantly at later stages, emphasizing the need for timely and accessible diagnostic tools. However, access to dermatological care remains limited for billions of people worldwide due to cost, geography, and time constraints. In this work, we present AutoLesion, an affordable and accessible artificial intelligence–based system for the preliminary assessment of skin lesion malignancy using a novel multimodal approach. Unlike prior methods that rely solely on dermoscopic or high-resolution imagery, AutoLesion integrates cell phone images with clinical and symptomatic metadata through a fine-tuned vision–language model (VLM). This joint utilization of visual and clinical information captures indications of malignant skin lesions that are often overlooked by image-only models. We further introduce a test-time compute strategy to improve prediction accuracy and reliability. Experimental evaluation on the ISIC (International Skin Imaging Collaboration) skin lesion archive demonstrates that the proposed approach outperforms dermoscopic image-only baselines, supporting the effectiveness of multimodal diagnosis from consumer-grade imagery. Even though these systems can improve early detection, especially in disadvantaged regions, these advantages are outweighed by practical and ethical concerns such as algorithmic bias, data privacy, and the need for human oversight. These considerations underscore the importance of responsible development and deployment of AI-assisted medical diagnostic tools.

Keywords: Artificial Intelligence (AI); Dermatology; Vision Language Models (VLM); Skin Cancer Detection; Medical image diagnosis; Multimodal Learning; Test-time Compute; Low-Rank Adaptation (LoRA)

INTRODUCTION

AI is rapidly advancing and is making a significant impact worldwide. From achievements such as Google's Self-Driving Car in 2009, to newer technologies like ChatGPT's ability to generate text, code, data, and vision, AI is becoming increasingly prevalent in daily life. These systems work because of major improvements in how computers learn from data and understand language, allowing them to think and act more like humans. Recently, deep learning (1), a part of machine learning involving neural networks with many layers, has boosted most of the progress. One of its most impactful applications is in the field of healthcare and medicine. Various AI techniques have found use in healthcare, medicine (2, 3) and medical image analysis, including

Corresponding author: Shreya Talukder, E-mail: shreyat2020@gmail.com.

Accepted June 17, 2026
https://doi.org/10.70251/HYJR2348.43646653

predictive models, deep learning (4), and Generative-AI LLMs, thereby greatly accelerating progress and improving medical diagnostic capabilities.

The goal of this research is to design an accessible, low-cost AI solution to diagnose the malignancy of a skin lesion using a novel combination of both visual and clinical information. Skin cancer, which can produce skin lesions as symptoms, is the most common type of cancer. Approximately, 1 in 5 Americans will develop skin cancer by age 70 (5), and more than 2 people die from skin cancer every hour in the U.S. Early-stage melanoma has a 99% five-year survival rate, but the chances of survival for late-stage melanoma drop dramatically, proving that early detection of malignant skin lesions is crucial. An AI model that can achieve malignancy detection early-on at home would shrink this number significantly because, according to the Skin Cancer Foundation, "early detection saves lives." People who don't visit the doctor for an appointment often cite "time constraints" as the reason, according to a study conducted by the National Library of Medicine. Furthermore, more than half of the world's population lacks coverage for essential health services or has limited access to treatment and medication (6). Our automated skin lesion diagnosis tool using high-resolution images (such as cell phone imagery), called AutoLesion, opens the ability for anyone with a cell phone and/or a camera to receive an initial skin lesion diagnostic, which is quick and effective, and works without access to specialized medical expertise. This AI model, able to conduct an initial diagnosis assessment of the skin lesion for malignancy at home within seconds, would be a good indicator to know if consultation by a medical professional is needed.

Without early detection or treatment, skin cancer can lead to a very low survival rate, spread throughout the body, and subsequently lead to limited options for treatments at later stages of the disease. The survival rate drops from 99% if it is detected before spreading to the lymph nodes, to 27% if it has spread to other organs. An AI model which can reliably diagnose a patient's skin lesion as benign or malignant using an image and essential clinical metadata, could significantly benefit underserved and rural communities, where dermatologists are often scarce. Current skin cancer detection requires expensive dermoscopic imaging and tools that are not available to everyone. By just needing a high-resolution image of the skin lesion (for example, on a mobile phone with a camera), and a short textual description of the individual's clinical attributes, this solution provides a cheaper, quicker, and more accessible diagnosis to the billions of individuals worldwide who have limited access to a medical professional. The novelty of AutoLesion is that it utilizes a combination of image and clinical attributes of the patient to estimate the malignancy of the lesion – both visual and symptomatic attributes jointly contain essential information for reliable diagnosis. The clinical and symptomatic patient attributes, available as text descriptions, provide essential information, resulting in more precise differential diagnosis. Most prior AI-based techniques for skin lesion classification use purely visual cues derived from mostly dermoscopic images (7, 8) or high-resolution images and largely ignore the clinical and symptomatic attributes of the patient, thereby missing out on critical information that is necessary for an accurate diagnosis. In our research, VLM that jointly utilizes image and text information about the patients' clinical conditions is fine-tuned for the task of skin lesion classification. Furthermore, an innovative test-time compute algorithm is used to improve the accuracy and reliability of the custom fine-tuned VLM. Results on the ISIC skin lesion archive demonstrate the effectiveness of the fine-tuned AI model for skin cancer diagnosis from cell phone images.

LITERATURE REVIEW

Prior research for automated skin lesion diagnosis includes primarily image-based classification from dermoscopic images or clinical images, including a support vector machine (SVM) machine learning and particle swarm optimization (PSO) principles for improving the performance of skin lesion in clinical diagnosis (9), MobileNet (10) with fused spatial channel attention mechanism, CNNs for image-based classification using dermoscopic images (7, 11), a custom image-based model which handles dermoscopic and clinical images (12), and a combination of pretrained image-based Deep Learning networks as feature extractors in conjunction with four shallow machine learning classifiers (8). All these prior techniques utilize the image of the skin lesion as the only source for classification. However, multiple clinical attributes such as patient history, demographics, physical attributes (including the patients age, gender, location, size of the skin lesion, duration of the lesion, personal and family history, exposure history, size and depth, quantity, arrangement, distribution, and others), offer detailed information that is useful for correct classification of the lesion by dermatologists (13), but are not used by prior

AI solutions. These features provide essential context that complements or, in some cases, replaces visual information, leading to a more precise diagnosis.

Convolutional Neural Networks (CNNs) (14) use an encoder-decoder architecture and have been the standard for computer vision tasks, particularly image segmentation. Image segmentation divides an image into smaller regions to analyze specific components in greater detail, such as identifying edges, objects, or textures. CNNs function by applying filters over input images, learning features through complex layers. Despite their effectiveness in visual tasks, CNNs can be limited in their flexibility. They are not well-suited for managing textual data, and their design does not naturally accommodate the sequential or contextual relationships present in language. For this reason, CNNs are unsuitable for tasks that require complex reasoning across multiple modalities like vision and text.

Large Language Models (LLMs) (15), on the other hand, operate using a decoder-only transformer architecture. These models process text by first tokenizing it, breaking it down into smaller units such as words or subwords. Then, it predicts the most likely next tokens in a sequence using a probability distribution, often modeled as a Gaussian to introduce stochasticity and variability into the output. This predictive and generative behavior is what makes LLMs generative AI. Since LLMs are built on the transformer architecture, they are very good at capturing long-range dependencies and context in text, making them very effective for a wide range of language understanding and generation.

Vision Transformers (ViT) (16) bring the transformer architecture from text into the area of images. Unlike CNNs, which process image data in a spatially localized manner, ViTs treat an image as a sequence of patches, and each section is turned into a list of numbers. These lists are then processed using self-attention mechanisms, similar to those used in LLMs. This allows the model to capture global context across an image, rather than just local features. Vision transformers have shown very strong performance in image classification and other vision tasks, taking advantage of how transformers can easily handle large amounts of data simultaneously. However, they are limited to visual inputs and do not incorporate language understanding.

Vision-Language Models (VLMs), such as the Qwen 2.5 VL models (17), extend the capabilities of ViTs by integrating both visual and textual modalities into a single model. Architecturally, VLMs usually consist of two parallel encoders, one for images and one for text, that project their respective inputs into a shared embedding space. Some models also include a cross-attention mechanism or use a unified transformer that jointly processes both modalities. This design allows VLMs to align visual information with language, allowing tasks like image captioning, visual question answering, and multimodal reasoning. Compared to ViTs, VLMs have a more flexible and powerful architecture for tasks that require understanding both images and text, making them increasingly valuable for applications in multimodal AI involving images and text.

METHODS AND MATERIALS

As discussed in the Literature Review Section, VLMs combine a vision encoder and a language model, and can handle simultaneous image and text data. VLMs can be fine-tuned for specific tasks, as discussed below. In this research, the Qwen 2.5 VL 7B Vision Language Model (VLM) base model was selected and fine-tuned for jointly analyzing the skin lesion images and the patient's text description of their clinical and symptomatic attributes to estimate the diagnosis of the skin lesions.

Dataset

The International Skin Imaging Collaboration (ISIC) Archive (18), which is a large, publicly available dataset of dermoscopic skin lesion images, was used for training and validating the AutoLesion solution. The archive has over 70,000 high-resolution dermoscopic, clinical, and cell-phone images, with a wide variety of benign and malignant lesions, including melanoma, basal cell carcinoma, actinic keratosis, and nevi. Many ISIC images also have patient-based clinical information metadata such as the patient's age, gender, location of the lesion, size and physical characteristics of the lesion, and family and personal medical history. Dermoscopic images require a specialized magnifying tool and are normally done at a specialized medical facility. Most of the non-dermoscopic images in the ISIC Archive are taken using either a digital camera, clinical camera, or smartphone. These images are high resolution, color images with the lesion generally centered in the frame, minimal shadows, and no flash. The high-resolution non-dermoscopic images in ISIC are analogous to those taken by a good quality camera from either a smartphone or iPad, that most people have access to at home. These images in ISIC are taken within a close range with variations in lighting conditions, similar to a close-up cell phone image capture at home. A subset of ~2200 of

these non-dermoscopic images with associated clinical information including patients' age, gender, and location of lesion were used for training and validation in our research. An 80-20 train-test split of this dataset was used for VLM fine-tuning and performance evaluation after tuning. The technique used for model fine-tuning is discussed next.

AI Model Fine-Tuning

Model Fine-tuning is a general approach to update a pretrained foundational model to accomplish specific tasks. It broadly involves building upon a pretrained model's pre-existing knowledge, and further modifying it on a smaller, domain-specific dataset to enhance the model's performance on specific tasks with reduced data and computational requirements (19). Full model fine-tuning updates all the model's parameters but needs a lot of computing power, and we opted not to use this method because of computing constraints. Full fine-tuning can take days to weeks depending on dataset size. In this research, a Low-Rank Adaptation (LoRA), technique was used which updates only small parts of the model to save memory. To enable model fine-tuning on constrained resources, a Quantized Low-Rank Adaptation (QLoRA) (20) was used in this research; QLoRA is a version of LoRA that quantizes the model's parameters using 4-bit quantization. This allows large models to be fine-tuned on less expensive consumer Graphics Processing Units (GPU). QLoRA uses significantly less memory than LoRA while generally preserving the performance of 16-bit full precision LoRA-based fine-tuning. Given our limited computational resources, we fine-tuned the Qwen 2.5 VL 7B model using QLoRA for skin-lesion diagnosis.

During fine-tuning, the VLM model was loaded in 4-bit mode to reduce memory use. The vision and language layers of the VLM were tuned using QLoRA. The Google Colab free-tier machine with an NVIDIA Tesla T4 GPU with 16GB GPU RAM (2,560 CUDA cores, 320 Tensor cores), 100GB disk space, and Intel Xeon 2.20GHz CPU were used for all fine-tuning experiments. A training batch size of 2 was used to limit memory usage. The hyperparameters employed during model fine-tuning include AdamW 8-bit optimization, gradient accumulation steps of 4, 40 Epochs, a learning rate of 2e-4, and a weight decay of 0.01 to reduce overfitting and improve generalization performance. The QLoRA configuration used a LoRA rank (r) of 16, LoRA alpha of 16, and LoRA dropout of 0. These hyperparameters were selected to balance model performance with the computational limitations of the Google Colab Tesla T4 GPU environment.

Additionally, the vision-language model received both clinical metadata and dermoscopic images in a conversational prompt format. The prompt template combined a fixed instruction, "Classify the image of the skin lesion as Malignant or Benign based on the image and patients provided clinical information in the text," with specific clinical metadata stored in the dataset text field, which included information such as patient age, gender, lesion location, and lesion dimensions. This combined text input was concatenated into a single instruction string and paired with the image of the skin lesion as multimodal input to the model. The assistant response for each training sample consisted of the classification label, enabling supervised fine-tuning of the classification task.

VLM Inference Methods: Test-Time Compute

Test-time compute (21) is a technique for a generative AI model to give more accurate answers by using more processing time during inference for each question or task. A survey of various test-time compute techniques is provided in (22). One simple but effective method is called Best-of-N Sampling (21). In Best-of-N, the model generates several responses, and we choose the one with the highest confidence, which could be a value generated by a reward algorithm. Majority Voting is one example of the Best-Of-N sampling technique, where the most common answer over N runs is selected as the final answer. For instance, this would be the same as solving a math problem many times in slightly different ways, and then we choose the answer it gives most often. This helps because small mistakes in decision making become less significant, as random errors are less likely to be repeated over further runs. Other test-time compute methods include beam search, and iterative self-refinement which all aim to improve a model's answers by changing how outputs are generated or selected. Beam search is a way for a model to pick better answers by keeping track of several possible options at once; instead of choosing just the most likely next word each time, it looks at the top few and builds multiple sentence paths at the same time. It only retains the best ones based on their overall likelihood. Beam search is commonly used in tasks like machine translation, where sequence and grammar matter. In VLMs, beam search can explore image captions or medical diagnoses paths more thoroughly than other methods. Iterative self-refinement involves generating an initial answer, then asking the model to

revise or improve it step by step. This reflects human problem-solving by repeatedly improving a draft. While all these methods can help, majority decisions are often better because they are the simplest and have a high chance of providing the correct output, as it relies on consistency. This makes majority decision more reliable, especially for tasks where high-confidence mistakes or wrong answers can have serious consequences. A model confidently giving a wrong cancer diagnosis is much worse than giving an uncertain one. That is why majority decision, which filters out outliers, is preferable.

RESULTS AND DISCUSSION

The ISIC archive data was filtered to down select only the subset of data comprised of cell-phone and clinical images and their corresponding clinical information. The clinical information used included patient age, gender, location of lesion, and approximate skin lesion dimensions. This data was divided to select 1000 malignant and 1000 benign samples. A 80-20 split was applied to create a training set of 1600 samples, comprised of 800 benign and 800 malignant data. The test set had 400 samples, with equal distribution between benign and malignant labels. The 1600 training samples were used to fine-tune the Qwen VLM model using QLoRA. The QLoRA fine-tuning technique and code (23) provided by Unsloth was used to fine-tune the model.

The performance of the fine-tuned VLM on the test set is shown in the confusion matrix in Table 1. The total accuracy of skin-lesion diagnosis was 78.5%. Table 1 discusses that the false-positive and false-negative rates for the VLM were 14% and 29% respectively on the test data. Ideally, these should be low to avoid false alarms to patients using AutoLesion.

Table 1. *Confusion Matrix on Test Set for Single Inference Run.*

		True Class	
		Benign	**Malignant**
ML	**Benign**	0.86	0.14
Label	**Malignant**	0.29	0.71

Next, the fine-tuned VLM was evaluated using the test-time compute paradigm. A majority-of-N voting scheme on N=9-inference runs for every test sample was conducted. While other test-time compute techniques exist as discussed earlier, the majority-of-N technique was simple to implement and practically feasible in this application. A simple-majority (M >=5) for N=9 inference test-time resulted in an overall accuracy of 90%. This corresponds to a relative improvement of 14.6% over a single VLM run discussed in Table 1. The confusion matrix for simple majority M=5 as shown in Table 2. Most importantly, benign cases were correctly detected 99% of the time. With a simple majority of 5, the false negative rate dropped to 1% (compared to 14% false negatives with single inference), while the false positive rate went down to 19% (compared to 29% false positives with the single inference method).

Table 2. *Confusion Matrix (Test Data) for Test-time Compute-Simple Majority.*

		True Class	
		Benign	**Malignant**
ML	**Benign**	0.99	0.01
Label	**Malignant**	0.19	0.81

While majority voting improved classification accuracy, it also increased inference cost because the model was executed multiple times for each test sample. For the N=9 majority voting strategy, inference computational requirements and latency increased significantly compared to a single inference run, since nine independent predictions must be generated before a final decision can be made. Although runtime measurements were not formally collected in this study, this additional computational overhead represents an important trade-off between diagnostic accuracy and deployment efficiency. In practical settings, the increased latency may be acceptable for non-emergency skin lesion screening applications, where improved diagnostic reliability is often more important than obtaining an immediate response.

To explore the performance of the test-time method further, a super majority of M=6 for both classes was evaluated. Super-majority implies that only data samples with high confidence for either benign or malignant classes are correctly classified; samples with low confidence are categorized as inconclusive. This results in fewer false positives (14%) and false negatives (1%), while lowering the precision for benign (96%) and malignant (73%) classes, as shown in Table 3.

Table 3. *Confusion Matrix (Test Data) for Test-time Compute-Simple Majority of 6.*

		True Class		
		Benign	**Malignant**	**Inconclusive**
ML	**Benign**	0.96	0.01	0.03
Label	**Malignant**	0.14	0.73	0.13

Supermajority test-time evaluations for higher majorities of M >= 7 and M>=8 were also conducted and the confusion matrices are shown below in Table 4 and Table 5. For a supermajority of 7, the false-positive reduced to 10% and false negative was 1%. Benign sample accuracy was 86% and Malignant accuracy dropped to 61%, and higher proportion of inconclusive results. Table 5 below shows results with a supermajority of 8 – meaning only highly confident samples are correctly classified, resulting in lower false positive (6%) and false negative (1%) outcomes, and lower accuracy precision for both classes. (64% for Benign, and 49% for Malignant) (Figure 1).

Table 4. *Confusion Matrix (Test Data) for Test-time Compute-Simple Majority of 7.*

		True Class		
		Benign	**Malignant**	**Inconclusive**
ML	**Benign**	0.86	0.01	0.13
Label	**Malignant**	0.1	0.61	0.29

Table 5. *Confusion Matrix (Test Data) for Test-time Compute-Simple Majority of 8.*

		True Class		
		Benign	**Malignant**	**Inconclusive**
ML	**Benign**	0.64	0.01	0.35
Label	**Malignant**	0.06	0.49	0.45

CONCLUSION

This research study involved the design of a novel VLM tuning solution to classify skin lesion from consumer-grade cell phone images combined with clinical metadata, and the results indicate that a fine-tuned VLM is a viable solution for skin-lesion classification based on this multimodal input strategy. The core purpose of this work was to explore whether a low-cost, accessible custom AI model fine-tuned using labeled skin lesion data could support early skin cancer screening without requiring specialized dermoscopic imaging or direct access to dermatologists.

In this research study, the AutoLesion approach combined both visual and clinical information and achieved improved diagnostic performance using test-time compute and majority decision methods, which show that VLMs can outperform models that make only one prediction at a time when applied to medical classification tasks. By supporting earlier and more accurate classification of benign and malignant lesions, AutoLesion has the potential to significantly improve patient outcomes and reduce delays in diagnosis. Beyond detection, the model is contributing to more personalized risk assessment by incorporating individual clinical histories and symptomatic information, while also accelerating research into improved diagnostic methods and treatment strategies. For patients and clinicians alike, this low-cost tool offers practical benefits such as remote preliminary screening, decision support, and increased

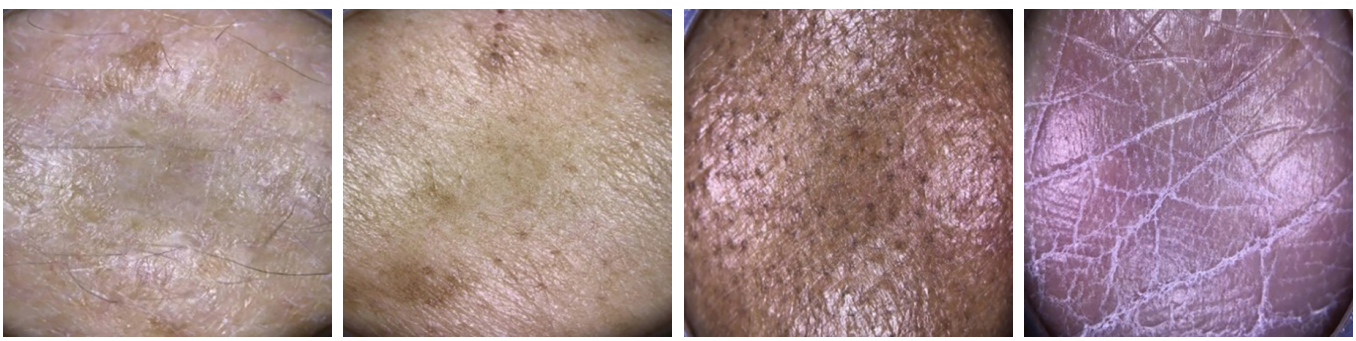

Figure 1. *Benign skin lesion samples that were incorrectly classified as Malignant in a single inference, but correctly labeled as Benign after Simple Majority Test-Time inference.*

access to dermatological expertise, particularly for individuals in underserved or resource-limited settings.

However, these advances with model driven diagnosis come with important responsibilities. As AI becomes more deeply integrated into dermatological care, it is essential to address issues of fairness and ensure that models perform reliably across diverse skin types and populations, without reinforcing existing healthcare disparities. Safeguarding patient privacy is equally critical, as sensitive medical images and clinical data must be collected and stored securely. Skin cancer is a heterogeneous disease influenced by numerous biological and environmental factors, and our models must be designed to reflect this complexity without oversimplification. Most importantly, AI should serve as a supportive tool that supplements rather than replaces clinical expertise and human judgment in skin cancer care. Additionally, current multimodal VLMs may struggle with variable lighting, camera quality, or incomplete metadata, which can reduce diagnostic confidence and reliability in real-world settings.

Future work should focus on improving the model so it can continue to learn from more diverse skin tones, camera types, fixing incorrect diagnosis, and reducing bias by using feedback from doctors or users. It will also be crucial to test the model on larger and more diverse groups of people to make sure it works fairly for everyone, preventing bias in the long run. In the long-term, the goal is to make AutoLesion easier to use in telehealth or online doctor visits and eventually run real clinical tests to make sure it is safe and helpful in the real world. Overall, these future efforts aim to develop AutoLesion from a research prototype into a reliable and effective tool that can support earlier skin cancer detection for people all over the world.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521: 436–444. https://doi.org/10.1038/nature14539
2. Angus D, Khera R, T L. AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA*. 2025; 334 (18): 1650–1664. 10.1001/jama.2025.18490
3. Xie Y, Zhai Y, Lu G. Evolution of artificial intelligence in healthcare: a 30-year bibliometric study. 2025; 11. 10.3389/fmed.2024.1505692
4. Esteva A, Robicquet A, Ramsundar B. A guide to deep learning in healthcare. *Nature Medicine*. 2019; 25: 24-29. https://doi.org/10.1038/s41591-018-0316-z
5. Skin Cancer Facts & Statistics, 2022. Available from: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/ (accessed on 2025-11-29).
6. WHO. World Bank and WHO: Half the world lacks access to essential health services, 100 million still pushed into extreme poverty because of health expenses. Tokyo:; 2017.
7. Nawaz K, Zanib A, Shabir I, Li J, Wang Y, Mahmood T, *et al*. Skin cancer detection using dermoscopic images with convolutional neural network. *Scientific Reports*. 2025; 15: 15(7252). https://doi.org/10.1038/s41598-025-91446-6
8. Shakya M, Patel R, Joshi S. A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification. *Scientific Reports*. 2025; 15: 15(4633). 10.1038/s41598-024-82241-w
9. Fei D, Almasiri O, Rafig A. Skin cancer detection using support vector machine learning classification based on particle swarm optimization capabilities. *Transactions on Machine Learning and Artificial Intelligence*. 2020; 8 (4): 1-13. https://doi.org/10.14738/tmlai.84.8415
10. Cheng H, Lian J, Jiao W. Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism. *Sci Rep* 2024; 14: 28850. https://doi.org/10.1038/s41598-024-80087-w
11. Esteva A, Kuprel B, Novoa R. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017; 542: 115–118. https://doi.org/10.1038/nature21056
12. Huang Y, Zhang Z, Ran X, Zhuang K, Ran Y. An Ingeniously Designed Skin Lesion Classification Model Across Clinical and Dermatoscopic Datasets. *Diagnostics (Basel)*. 2025 August; 15 (16). https://doi.org/10.3390/diagnostics15162011
13. Cleveland Clinic (Skin Lesions), 2015. Available from: https://my.clevelandclinic.org/health/diseases/24296-skin-lesions (accessed on 2025-09-05).
14. Huang G, Liu Z, Maaten Lvd, Weinberger K. Densely Connected Convolutional Networks (CVPR 2017, Best Paper Award). In IEEE CVPR; 2017. https://doi.org/10.1109/CVPR.2017.243
15. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, *et al*. Language models are few-shot learners. In NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020. https://doi.org/10.48550/arXiv.2005.14165

16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, *et al*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR; 2021. https://doi.org/10.48550/arXiv.2010.11929
17. Bai S, Chen K, Liu X, Wang J, W G, Song S, *et al*. Qwen2. 5-VL Technical Report. 2025. https://doi.org/10.48550/arXiv.2502.13923
18. ISIC. The International Skin Imaging Collaboration, 2016. Available from: https://www.isic-archive.com/ (accessed on 2025-09-05).
19. Parthasarathy VB, Zafar A, Khan A, Shahid A. The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. Dublin, Ireland:; 2024. https://doi.org/10.52202/075280-0441
20. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient fine-tuning of quantized LLMs. In NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023; p.10088 - 10115. https://doi.org/10.48550/arXiv.2305.14314
21. Xuezhi W, Wei J, Schuurmans D, Le Q, Chi E, Narang S, *et al*. Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations (ICLR) 2023; 2023. https://doi.org/10.48550/arXiv.2203.11171
22. Ji Y, Li J, Xiang Y, Ye H, Wu K, Yao K, *et al*. A Survey of Test-Time Compute: From Intuitive Inference to Deliberate Reasoning. 2025. https://doi.org/10.48550/arXiv.2501.02497
23. Unsloth. Unsloth Notebooks, 2025. Available from: https://colab.research.google.com/github/unslothai/notebooks/blob/main/nb/Qwen2.5_VL_(7B)-Vision.ipynb (accessed on 2026-01-04).