2025 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) | 979-8-3315-0266-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/CIBCB66090.2025.11177070

# Evaluating the Feasibility of Vision-Language Models in Skin Cancer Detection: A Comparative Study with CNNs and Vision Transformers

Yu-Hxiang Chen*, Ting-Ting Chang†, Yao-Zhi Xue*, Wei-Hsiang Sung*, and Chia-Yu Lin*
* *Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan*
† *Department of Information Management, National Central University, Taoyuan, Taiwan*
Corresponding Author: Chia-Yu Lin (sallylin0121@ncu.edu.tw)

***Abstract*—This study presents a comprehensive evaluation of vision-language models (VLMs) for skin lesion classification, comparing them with conventional convolutional neural networks (CNNs) and Vision Transformer (ViT) architectures. Using the HAM10000 dataset, we evaluated six models in four key dimensions: classification accuracy, robustness to visual perturbations, zero-shot generalization to unseen lesion types, and the quality of semantic explanations. Although traditional vision-only models achieve higher accuracy in clean images, their performance degrades significantly under various types of image distortion. In contrast, VLMs—particularly Qwen2.5—demonstrate stronger robustness and generate more coherent and clinically relevant explanations. However, they still fail to achieve overall classification performance and exhibit limited generalization in zero-shot settings. These findings highlight the trade-offs between task-specific accuracy and multimodal adaptability, offering practical insights into the current capabilities and limitations of VLMs in dermatology-focused artificial intelligence applications.**

***Index Terms*—Vision-Language Models, CNNs, Vision Transformers, Skin cancer detection**

## I. Introduction

Skin cancer is among the most common malignancies worldwide, with melanoma being the deadliest due to its rapid metastasis. The early and accurate diagnosis of skin lesions is critical to improving outcomes. Deep learning, particularly convolutional neural networks (CNNs), has advanced automated skin lesion classification by enabling end-to-end feature learning from dermoscopic images. However, CNNs often struggle with generalization and interpretability, especially when faced with diverse skin tones, lighting conditions, and imaging artifacts.

To address these limitations, transformer-based models, such as Vision Transformers (ViTs), have been introduced. Using self-attention mechanisms, ViTs capture global context more effectively than CNNs and have demonstrated promising results in medical imaging. However, their high data requirements, computational costs, and vulnerability to visual degradation remain challenges.

Recent advances in Vision-Language Models (VLMs) offer a new paradigm that jointly leverages visual and textual information. Pre-trained on large-scale multimodal datasets, VLMs can perform various vision tasks with minimal supervision, support zero-shot inferences, and generate human-readable explanations. These properties suggest a strong potential to improve the robustness, generalization, and transparency of clinical AI systems.

This study conducts a comprehensive evaluation of VLMs in skin lesion classification, comparing them with CNN and ViT baselines. We evaluated the models in four dimensions: classification accuracy, robustness to visual perturbations, zero-shot generalization to unseen classes, and semantic explanation quality. By analyzing performance across these axes, we aim to clarify the capabilities and limitations of current VLMs and provide guidance for their application in dermatology decision support systems.

## II. Related Work

Various models have been developed for skin cancer detection. This section reviews their evolution and limitations, beginning with CNNs and ResNets, followed by Vision Transformers (ViTs) and culminating with Vision-Language Models (VLMs).

### A. CNNs and ResNets in Skin Cancer Analysis

Convolutional Neural Networks (CNNs) have served as a foundational framework in computer vision, thanks to their capacity to learn hierarchical spatial representations through layered convolutions and pooling operations. By applying fixed-size kernels locally, CNNs excel at detecting low- to mid-level patterns such as edges, textures, and object shapes—an essential strength for tasks like skin lesion classification.

Pioneering models such as LeNet-5 [1], AlexNet [2], and VGGNet [3] laid the groundwork for deep learning in visual recognition. However, increasing network depth introduced significant challenges, including vanishing gradients and performance degradation. To mitigate these issues, He et al. introduced Residual Networks (ResNets) [4], incorporating identity shortcuts to create residual blocks. These shortcuts enhance gradient propagation, enabling the successful training of much deeper architectures. ResNet models like ResNet-50 and ResNet-101 have since become standard backbones in medical imaging, particularly under transfer learning settings.

In dermatological applications, CNNs—especially ResNet variants—have shown strong performance on dermoscopic datasets, such as those from the ISIC archive [5], for classifying lesion types like melanoma, basal cell carcinoma, and

979-8-3315-0266-9/25/$31.00 ©2025 IEEE

benign nevi. Notably, fine-tuned ResNet-50 models on ISIC 2018 data have achieved impressive classification accuracy and AUC scores [6]. Several studies have even demonstrated that deep residual networks can rival or outperform dermatologists in diagnostic accuracy [7], [8]. Leveraging pretrained CNNs through transfer learning has become a widely adopted solution to compensate for the limited availability of labeled medical images.

Nevertheless, CNN-based approaches still encounter obstacles in clinical practice. Their reliability can diminish under variable image conditions caused by lighting inconsistencies, noise, or resolution loss [9]. Moreover, training data biases can hinder generalization across diverse skin tones and demographic groups [10]. Finally, their opaque decision-making process remains a barrier to clinical acceptance. Although interpretability techniques like Grad-CAM offer partial insights, they often fall short of providing the level of transparency necessary for trustworthy medical decision-making [11].

### *B. Vision Transformers in Dermatology*

To address the limitations of local receptive fields in convolutional neural networks (CNNs), Dosovitskiy et al. proposed the Vision Transformer (ViT) [12], which employs self-attention over image patches to capture global context. Unlike CNNs, ViTs represent images as sequences of fixed-size patches and process them using transformer encoders, allowing the model to capture long-range dependencies from the early layers.

ViTs have shown strong performance in dermatological image analysis. Dahmani et al. [13] reported 94% and 97.2% accuracy on the HAM10000 and benign-vs-malignant datasets using ViT-B16. Similarly, Shajimon et al. [14] developed a compact ViT with focal loss and dropout that outperformed Inception-ResNetV2 on ISIC 2017. These results highlight ViT's strength in capturing global structures and symmetry in dermoscopic images.

However, ViTs face notable challenges. They require large labeled datasets for stable convergence, limiting their use in data-scarce domains. While fine-tuning and augmentation have been employed [13], overfitting remains an issue. ViTs are also sensitive to class imbalance [15] and generalize poorly across variations in skin tone, lighting, and acquisition devices [16]. Their high computational demands hinder deployment, and although attention maps enhance interpretability, they often fall short of clinical transparency.

### *C. Vision-Language Models for Interpretable Medical Imaging*

Vision-Language Models (VLMs) have emerged as a solution to the limitations of unimodal vision systems, particularly in interpretability and human-AI interaction. By jointly encoding visual and textual inputs, VLMs enable semantic reasoning and natural language-based inference, making them especially relevant for high-stakes applications like medical imaging.

For instance, Florence-2 [17] adopted a unified encoder-decoder architecture with prompt-based modeling and achieved strong performance in image captioning, detection, and segmentation. Qwen-VL [18] integrated a ViT-based encoder with a large language model, using cross-attention and 2D positional encoding to retain spatial structure. Phi-3.5-Vision [19] combined a CLIP-based visual encoder with a lightweight decoder, supporting multi-image inputs, document understanding, and structured vision tasks via multimodal pretraining.

In the medical domain, models such as CLIP, LLaVA, and BiomedCLIP have been adapted for diagnostic reasoning across modalities, including X-ray, MRI, and histopathology [20]. These models leverage large-scale vision-language pretraining to enable zero-shot and prompt-based learning for clinical queries and visual understanding.

To improve generalization, recent efforts incorporate contrastive pretraining (e.g., CLIP, ConVIRT) [21], medical knowledge integration through ontologies and graphs [22], and parameter-efficient fine-tuning techniques such as LoRA, prefix tuning, and instruction tuning [23], [24]. Additionally, self-supervised learning enables scalable training on unlabeled datasets, which is particularly valuable in data-scarce clinical environments.

Interpretability remains a key concern. Techniques like LIME have been used to visualize model rationale and improve clinician trust [21]. Architectures such as the Mutual Attention Transformer (MAT) [25] and Concept Bottleneck Models (CBMs) [26] enhanced explainability by aligning predictions with domain-specific concepts.

Among clinical specialties, dermatology presents unique challenges and opportunities for VLMs due to its reliance on visual diagnosis, variation in skin presentations, and the critical need for interpretable and trustworthy AI support.

### *D. Summary*

Over the past decade, deep learning has revolutionized skin cancer diagnosis through successive innovations—from CNNs to Vision Transformers (ViTs), and more recently, Vision-Language Models (VLMs). While CNNs excel at localized feature extraction and ViTs capture global context, VLMs integrate visual perception with natural language reasoning.

Despite this progress, challenges remain in generalization, robustness, and interpretability. To address these gaps, this study investigates the application of VLMs to skin lesion classification by integrating dermoscopic images with structured clinical metadata.

## III. Methodology

### *A. Dataset*

This work utilizes the HAM10000 dataset [27], a publicly accessible repository containing 10,015 dermoscopic images, each depicting a single skin lesion. Expertly curated by dermatologists, HAM10000 has become a standard benchmark in studies focused on automated skin cancer diagnosis and dermoscopic image analysis.

Each image is labeled with one of seven diagnostic categories: actinic keratoses (AKIEC), basal cell carcinoma

(BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanocytic nevi (NV), vascular lesions (VASC), and melanoma (MEL).

The dataset includes a structured metadata file (HAM10000_metadata.csv) containing fields such as lesion ID, image filename, diagnosis label, diagnostic method, patient age and sex, and anatomical site. This metadata supports demographic and anatomical analyses and serves as the basis for generating textual prompts in the vision-language model (VLM) pipeline.

### *B. Data preprocessing*

To ensure consistent model training and evaluation, the dataset is split into training, validation, and testing sets in a 7:2:1 ratio. Four diagnostic categories (BKL, NV, MEL, and BCC) are selected, with an equal number of samples per class—210 for training, 60 for validation, and 30 for testing—to prevent class imbalance.

All images are resized and normalized based on the input requirements of each model. For vision-language model (VLM) evaluation, each data sample is paired with a natural language description generated from its metadata, including age, sex, lesion location, and diagnosis. These descriptions are further enriched with domain-specific definitions to provide relevant diagnostic context. This multimodal setup enables the models to utilize both visual and textual information.

A standardized system prompt is included with each entry, specifying the model's role, the classification task, category definitions, and the expected output format. Image paths are constructed according to dataset partitions and file-naming conventions. The final dataset, which contains prompts, image paths, ground-truth labels, and simulated rationale fields, is exported as a CSV file for downstream inference and evaluation.

### *C. CNN-based Models*

We implement two convolutional neural network models to establish image-only baselines for skin-cancer classification.

*1) Naive CNN:* A compact convolutional neural network (CNN) is developed for processing RGB dermoscopic images resized to 100 × 75 pixels. The model comprises four convolutional layers with 16, 32, 64, and 64 filters, respectively. Each of these layers is followed by a ReLU activation and a max-pooling operation. All convolutional layers utilize 3 × 3 kernels with "same" padding to maintain the spatial dimensions. After feature extraction, the resulting maps are flattened and passed through two dense layers with 32 and 16 units, respectively, before reaching a softmax output layer that performs four-class classification. The network is trained for 15 epochs using the Adam optimizer and sparse categorical cross-entropy loss, with a batch size of 32. A model checkpoint mechanism is used to save the version achieving the highest validation accuracy.

*2) ResNet Models:* The second model leverages the ResNet-50 architecture, which has been pretrained on the ImageNet dataset. Its final fully connected layer is adapted to produce outputs corresponding to four diagnostic classes. Before training, input images are normalized. The model is trained for 10 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-4}$ and employs cross-entropy loss. Validation accuracy is tracked after each epoch to evaluate model performance.

These CNN-based architectures function as baseline models for comparing the classification accuracy and robustness of transformer-based and vision-language approaches.

### *D. Transformer-based Model*

For the transformer-based approach, the Vision Transformer (ViT-Base) model pretrained on the ImageNet-21k dataset is adopted. All input images are normalized and resized to 224 × 224 pixels using the corresponding pretrained image processor. A classification head is added to the pretrained backbone to produce four output classes. Label mappings are explicitly defined to align numeric indices with class names.

Training is conducted for 4 epochs with a batch size of 16, using mixed-precision floating-point acceleration. The AdamW optimizer is applied with a learning rate of $2 \times 10^{-4}$. Model evaluation and checkpointing are performed at regular intervals, with the best-performing model selected based on validation accuracy. Accuracy serves as the primary metric to monitor training progress.

A custom data collator is implemented to aggregate pixel values and labels during training, enabling efficient batching and data loading.

This transformer-based model functions as a high-capacity image-only baseline, allowing direct comparison with the convolutional and vision-language model architectures.

### *E. VLM-based Models*

To investigate the potential of vision-language models (VLMs) in skin cancer classification, we fine-tune three state-of-the-art multimodal models: Florence-2 Base, Qwen2.5-VL 7B, and Phi-3.5-Vision. Each model is trained on structured inputs that combined dermoscopic images with textual metadata, enabling both image classification and explanation generation. All three models are fine-tuned using 4-bit quantization to reduce memory usage and computational costs.

*1) Florence-2 Base Fine-tuning:* Florence-2 Base, developed by Microsoft, is fine-tuned using image-text pairs. Each input sample consists of a dermoscopic image paired with a prompt that integrates metadata—such as patient age, sex, and lesion location—followed by a classification query. During training, the model's vision encoder remains frozen while the language model parameters are updated.

The model is trained for 20 epochs using the AdamW optimizer with a learning rate of $1 \times 10^{-6}$. Token-level accuracy, recall, and F1-score are used to evaluate performance at each epoch.

*2) Qwen2.5-VL 7B with LoRA:* Qwen2.5-VL 7B is a multimodal instruction-following model built on a 7-billion-parameter language backbone. The model is quantized to 4-bit precision and fine-tuned using Low-Rank Adaptation (LoRA)

applied to selected projection layers within the language model.

Training inputs follow a dialogue-style format that includes system instructions, user queries, and assistant responses, with images embedded within the prompt. The model is trained for 10 epochs using the AdamW optimizer with gradient accumulation and periodic checkpointing. Evaluation focuses on classification accuracy and the coherence of the generated responses.

*3) Phi-3.5-Vision with LoRA:* Phi-3.5-Vision was fine-tuned using 4-bit NormalFloat (nf4) quantization alongside bfloat16 mixed-precision to optimize training efficiency and memory usage. Low-Rank Adaptation (LoRA) was applied specifically to the query and value projection layers, configured with a rank of 16, a scaling factor of 32, and a dropout rate of 0.05.

Each training sample comprised a system prompt, a user query, and a dermoscopic image of a skin lesion. Textual inputs were tokenized to a maximum of 512 tokens, while images were processed through the model's built-in vision encoder. The model underwent training for 15 epochs using the Adam optimizer, with a learning rate of $5 \times 10^{-5}$, a batch size of 1, and gradient accumulation set to 4. Validation metrics were evaluated every 100 steps to track model performance.

*4) Model Objective:* All three VLMs were trained on multimodal inputs that integrated dermoscopic images and structured clinical metadata. In addition to lesion classification, these models were designed to generate natural language explanations, supporting the downstream evaluation of semantic reasoning and interpretability.

### *F. Performance Evaluation*

To comprehensively assess model performance, four evaluation dimensions are considered: classification accuracy, robustness to visual perturbations, generalization in vision-language models, and the quality of semantic reasoning.

*1) Classification Metrics:* Model performance is measured using accuracy, recall, and F1-score. In clinical applications, recall is particularly important, as false negatives may delay diagnosis and treatment. Therefore, sensitivity is emphasized alongside accuracy and the balanced F1-score.

*2) Robustness to Visual Perturbations:* To evaluate model robustness under degraded visual conditions, three types of perturbations are applied: Gaussian blur, resolution downsampling, and gamma correction to simulate overexposure. These tests assess how well the models maintain performance under non-ideal input conditions.

*3) Generalization in Vision-Language Models:* To evaluate the zero-shot generalization capabilities of the vision-language models, we introduced three diagnostic categories that were excluded from training. These novel classes were presented using natural language prompts, enabling the models to infer their meanings without direct exposure to labeled examples. This setup allowed us to assess the models' ability to classify previously unseen conditions using only descriptive text.

TABLE I: Classification metrics of different models

| Model | Accuracy | Recall | F1-score |
|---|---|---|---|
| Naive CNN | 0.56 | 0.57 | 0.56 |
| ResNet50 | **0.72** | **0.72** | **0.73** |
| ViT Transformer | 0.69 | 0.69 | 0.68 |
| Florence-2 | 0.28 | 0.26 | 0.10 |
| Qwen2.5 | 0.62 | 0.62 | 0.60 |
| Phi3-5-vision | 0.4 | 0.43 | 0.29 |

*4) Semantic Reasoning and Explanation Quality:* The quality of model-generated explanations is assessed using both automatic and subjective metrics. BLEU and ROUGE scores measure syntactic overlap, while BERTScore evaluates semantic similarity. Additionally, a large language model is used as an evaluator to provide qualitative feedback on the clarity and relevance of the generated explanations.

## IV. Experiments

### *A. Classification Performance Overview*

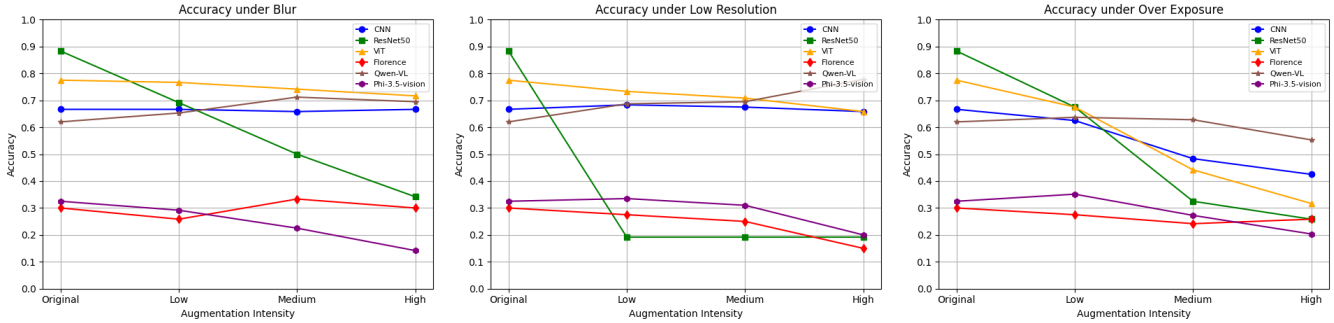
Table I presented the classification performance of all models fine-tuned on the target dataset. Among these, conventional vision-only models outperformed their multimodal counterparts. ResNet50 achieved the highest accuracy of 0.72, followed by the ViT Transformer at 0.69 and the Naive CNN at 0.56. These results underscored the strong task-specific learning capabilities of convolutional and transformer-based architectures in structured medical image classification.

In contrast, vision-language models (VLMs) demonstrated lower classification performance. Qwen2.5 yielded moderate results with an F1-score of 0.60, while Phi3.5-Vision and Florence-2 performed considerably worse, particularly Florence-2, which reached an F1-score of only 0.10. Further analysis revealed that both Florence-2 and Phi3.5-Vision frequently defaulted to predicting a limited subset of dominant classes, leading to imbalanced per-class metrics and reduced macro-level accuracy. These findings suggested that, despite fine-tuning, general-purpose VLMs lacked the domain-specific robustness required for accurate medical image classification.

### *B. Robustness Under Visual Perturbations*

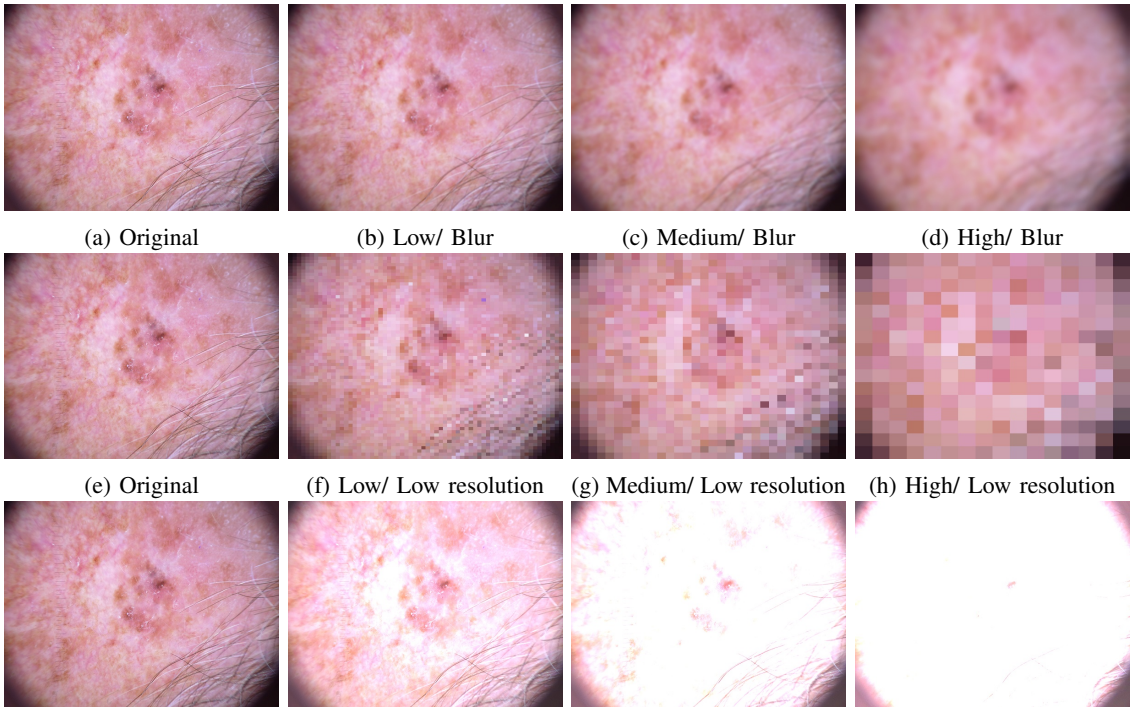
To assess model robustness, we evaluated six models under three types of image perturbations—blur, low resolution, and overexposure—each applied at multiple intensity levels. Figure 1 summarizes the performance trends, and Figure 2 illustrates an example of a skin lesion image subjected to these perturbations.

*1) Blur:* Under blur perturbation, ResNet50 achieved the highest accuracy on clean images (0.88) but suffered the steepest performance degradation, dropping to 0.35 under severe blur. This indicated a strong reliance on sharp visual features. In contrast, Qwen2.5 demonstrated notable robustness, with performance slightly improving at moderate blur levels. ViT




(a) Blur (b) Low resolution (c) Over exposure

Fig. 1: This figure illustrates the classification accuracy of six fine-tuned models under three types of image perturbations (blur, low resolution, and overexposure), each applied at three intensity levels: low, medium, and high.



(a) Original (b) Low/ Blur (c) Medium/ Blur (d) High/ Blur

(e) Original (f) Low/ Low resolution (g) Medium/ Low resolution (h) High/ Low resolution

(i) Original (j) Low/ Over exposure (k) Medium/ Over exposure (l) High/ Over exposure

Fig. 2: The image presents a representative skin cancer example from the HAM10000 dataset [27], subjected to three types of image perturbations: Gaussian blur, low resolution, and overexposure.

showed a moderate decline, while Florence-2 and Phi3.5-Vision consistently performed poorly, maintaining accuracy below 0.35 across all blur levels.

*2) Low Resolution:* Resolution reduction caused a substantial decline in ResNet50 performance, falling to 0.20 under low-resolution conditions. ViT and CNN maintained relatively stable accuracy, with ViT showing the highest robustness among the vision-only models. Qwen2.5 again demonstrated consistent performance, ranging from 0.62 to 0.65, suggesting resilience to low-frequency inputs—likely a result of its vision-language alignment training. Florence-2 exhibited a steady decline, reaching 0.15, while Phi3.5-Vision showed a downward trend following an initial plateau.

*3) Overexposure:* Overexposure uniformly degraded model performance. ResNet50 and ViT experienced the most significant drops, reaching 0.25 and 0.31, respectively. Qwen2.5 maintained the most stable accuracy, declining only slightly from 0.62 to 0.55, again reflecting its robustness under visually adverse conditions. Florence-2 and Phi3.5-Vision exhibited consistently low performance throughout.

These results revealed a key trade-off: while vision-only models achieved higher accuracy on clean data, they were more vulnerable to visual perturbations. In contrast, vision-language models—particularly Qwen2.5—exhibited greater robustness, likely due to their multimodal training that emphasizes semantic understanding over reliance on fine-grained

TABLE II: Classification Performance under Generalization Scenario

| Model | Accuracy | Recall | F1-score |
|---|---|---|---|
| Florence-2 | 0 | 0 | 0 |
| Qwen2.5 | **0.49** | **0.49** | **0.39** |
| Phi3.5 | 0.36 | 0.36 | 0.23 |

TABLE III: Quantitative Evaluation of Semantic Explanation Quality

| Model | BLEU | ROUGE-L | BERTScore | LLM Criticizer |
|---|---|---|---|---|
| Florence-2 | 0.23 | 0.43 | 0.91 | 1.80 |
| Qwen2.5 | **0.59** | **0.71** | **0.95** | **2.49** |
| Phi3.5 | 0 | 0.17 | 0.85 | 1.71 |

visual detail.

### C. Generalization Scenarios

Table II summarizes the performance of the VLMs in the three diagnostic categories excluded from training: AKIEC, VASC, and DF.

Florence-2 failed to generalize, consistently predicting the original training classes despite the prompt constraints, resulting in zero accuracy and recall. Phi3.5 followed the label constraints but predicted almost exclusively the AKIEC class, yielding an F1-score of 0.23. Qwen2.5 outperformed both models, correctly identifying the two novel classes and achieving an F1-score of 0.39, although it failed to detect the DF class. Overall, these findings highlight the limited zero-shot generalization capabilities of current VLMs in high-difficulty clinical conditions.

### D. Evaluation of Semantic Reasoning and Explanations

To evaluate the semantic reasoning capabilities of VLMs, we assessed their generated explanations using BLEU, ROUGE-L, BERTScore, and a subjective LLM-based critic score (Table III).

Qwen2.5 achieved the highest scores across all evaluation metrics—BLEU: 0.59, ROUGE-L: 0.71, BERTScore: 0.95, and LLM Critic: 2.49. Its responses were fluent, logically structured, and frequently incorporated domain-specific evidence. Florence-2 produced moderately coherent outputs (BERTScore: 0.91), but its lower BLEU and ROUGE scores suggested weaker organization and formatting. Phi-3.5 performed the poorest, particularly in BLEU and ROUGE-L, likely due to its minimal and unstructured responses, despite achieving a moderate BERTScore of 0.85.

While all models maintained basic fluency, only Qwen2.5 consistently generated explanations with meaningful content and clear logical reasoning. These findings underscored the limitations of current VLMs in handling complex semantic tasks, despite their general proficiency in natural language generation.

## V. Limitations and Future Work

Despite promising results, this study has several limitations that may impact the generalizability and interpretability of the findings. First, the evaluation was based solely on the HAM10000 dataset, which may not reflect the full visual and semantic diversity of real-world clinical settings. Some models may have benefited from dataset biases, such as the overrepresentation of certain lesion types. Additionally, the training set was limited to 210 samples per class—insufficient for large-scale vision-language models (VLMs) to converge effectively, especially given the high intra-class variability of skin lesions.

Second, while we used established metrics (BLEU, ROUGE-L, BERTScore, and LLM-based evaluations), each has limitations in the medical domain. Overlap-based metrics penalize valid paraphrasing and fail to capture semantic similarity. BERTScore, though better, often conflates medical terms, leading to uniformly high scores that mask clinically important differences. LLM-based evaluations, while helpful for assessing explanation quality, are sensitive to prompt design and model choice, which affects reproducibility and objectivity.

Future work should focus on developing a unified benchmarking framework for VLMs in dermatology. This includes creating a standardized, diverse dataset that covers a wide range of lesion types, skin tones, and imaging conditions, along with semantically rich class descriptions. Evaluation protocols should adopt clinically informed metrics that go beyond token overlap and better reflect diagnostic relevance.

Lastly, our robustness analysis used only four discrete perturbation levels, which may overlook critical thresholds where model performance deteriorates. Finer-grained perturbation spectra could offer more detailed insights. Moreover, normalizing model capacity across architectures would enable fairer comparisons, isolating architectural effects from model scale and supporting the principled design of multimodal diagnostic systems.

## VI. Conclusion

This study presented a comprehensive evaluation of vision-language models (VLMs) for skin cancer classification, benchmarking them against conventional CNN and transformer-based models. By integrating dermoscopic images with structured metadata, we assessed performance across multiple dimensions, including classification accuracy, robustness to visual perturbations, generalization to unseen categories, and semantic reasoning quality.

While traditional vision-only models such as ResNet50 and ViT achieved higher accuracy on clean data, they showed limited robustness under image degradation. In contrast, VLMs—particularly Qwen2.5—demonstrated greater resilience and produced clinically relevant explanations. However, their overall classification accuracy remained lower, and generalization to novel classes was still limited.

These findings highlight both the promise and the current limitations of multimodal models in medical imaging. The

ability of VLMs to incorporate textual context and generate interpretable outputs offers a compelling path toward more transparent and adaptable AI systems for dermatology.

## ACKNOWLEDGEMENTS

This work is sponsored by the National Science and Technology Council (NSTC) under the project NSTC 113-2222-E-008-002.

## REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] N. Codella, V. Rotemberg, P. Tschandl *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2018.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, p. 180161, 2019.

[7] T. J. e. a. Brinker, "Convolutional neural networks are superior to dermatologists in melanoma image classification," *European Journal of Cancer*, vol. 119, pp. 11–17, 2019.

[8] H. A. e. a. Haenssle, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[9] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. Rademaker, "Evaluating deep learning for skin disease classification," *Medical Journal of Australia*, vol. 212, no. 11, pp. 528–532, 2020.

[10] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Y. Zou, "Disparities in dermatology ai performance on a diverse dataset," *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, 2021.

[11] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[12] A. e. a. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[13] M. G. Dahmani, M. Tarhouni, and S. Zidi, "Vision transformers (vit) for enhanced skin cancer classification," in *2024 IEEE International Conference on Artificial Intelligence and Green Energy (ICAIGE)*, 2024.

[14] G. M. Shajimon, I. Ufumaka, and H. Raza, "An improved vision-transformer network for skin cancer classification," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 2213–2216.

[15] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2980–2988.

[16] Y. Gulzar and S. A. Khan, "Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study," *Applied Sciences*, vol. 12, no. 12, p. 5990, 2022.

[17] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," 2023. [Online]. Available: https://arxiv.org/abs/2311.06242

[18] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023. [Online]. Available: https://arxiv.org/abs/2308.12966

[19] M. A. et al., "Phi-3 technical report: A highly capable language model locally on your phone," 2024. [Online]. Available: https://arxiv.org/abs/2404.14219

[20] M. Van, P. Verma, and X. Wu, "On large visual language models for medical imaging analysis: An empirical study," in *IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2024, pp. 172–176.

[21] Z. K. Butt, A. K. Butt, T. Zia, and M. K. Butt, "Medical visual question answering using contrastive language-image pre-training," in *International Conference on Frontiers of Information Technology (FIT)*, 2024, pp. 1–6.

[22] X. Wei, Z. Vagena, C. Kurtz, and F. Cloppet, "Integrating expert knowledge with vision-language model for medical image retrieval," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024.

[23] S. Wu, B. Yang, Z. Ye, H. Wang, H. Zheng, and T. Zhang, "Maken: Improving medical report generation with adapter tuning and knowledge enhancement in vision-language foundation models," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[24] J. Luo, H. Yu, C. Tan, and H. Yu, "Enhanced qwen-vl 7b model via instruction finetuning on chinese medical dataset," in *2024 5th International Conference on Computer Engineering and Application (ICCEA)*, 2024, pp. 526–530.

[25] L. Zhou and Y. Luo, "Deep features fusion with mutual attention transformer for skin lesion diagnosis," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 3797–3801.

[26] C. Patrício, L. F. Teixeira, and J. C. Neves, "Towards concept-based interpretability of skin lesion diagnosis using vision-language models," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018. [Online]. Available: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000