Explainable SLM-Guided Vision–Language Model for Multi-class Skin Lesion Recognition

J. Jeffin Gracewell, J. Yuvaraj, and S. Vikhram(✉)

Department of Electronics and Communication Engineering, Saveetha Engineering College, Chennai, India
{jeffingracewellj,vikhrams}@saveetha.ac.in

Abstract. We present a compact and explainable multimodal frame-work—the Small Language Model–Guided Vision–Language Integration (SLM-VLI)—for multi-class skin lesion recognition. The approach cou-ples a lightweight visual encoder (MobileNetV3) with a small language model (SLM) to achieve parameter-efficient multimodal fusion and in-terpretable outputs. A cross-attention fusion block aligns SLM-derived semantic embeddings with visual features and a hybrid explainability module combines Grad-CAM saliency with SLM-generated tex-tual ex-planations. We evaluate our approach on the DermaMNIST benchmark, reporting classification accuracy, macro-F1, per-class metrics, calibra-tion, and computational cost comparisons against lightweight baselines and larger VLMs. Our SLM-VLI achieves competitive predictive perfor-mance while using substan-tially fewer parameters and providing clini-cally meaningful explanations. We also provide extensive ablation stud-ies, analysis of failure modes, and an open-source implementation to support reproducibility.

Keywords: Skin lesion classification · Small language models · Vision–language fusion · Explainable AI · DermaMNIST

1 Introduction

Skin malignancies represent among the most frequently diagnosed cancers glob-ally, and their incidence has continued to increase in recent years.Early detection significantly improves survival rates, but access to experienced dermatologists remains limited in many healthcare systems. Deep learning–based computer-aided diagnostic systems have recently demonstrated strong potential for au-tomated dermoscopic image analysis, but practical deployment requires models that achieve strong predictive performance while remaining computationally ef-ficient and interpretable [11, 13].

Vision–Language Models (VLMs) provide the ability to align visual patterns with natural language, supporting both classification and explainable text out-puts [2, 3]. However, most state-of-the-art VLMs rely on large language encoders that are impractical for mobile and edge deployment due to heavy compute and memory requirements [5]. Meanwhile, Small Language Models (SLMs) such as DistilBERT and MiniLM offer

A. E. Hassanien et al. (Eds.): ICICC 2026, LNNS 2045, pp. 611–622, 2026.
https://doi.org/10.1007/978-3-032-30310-3_49

compact alternatives with significant parameter savings while retaining strong semantic encoding capacity [6, 7].

This work introduces the **SLM-Guided Vision–Language Integration** (SLM-VLI) framework. SLM-VLI uses MobileNetV3 as an efficient visual back-bone, a compact SLM for semantic prompt encoding, and a cross-attention fusion block to combine modalities. To support clinical acceptance, SLM-VLI produces dual-mode explanations: Grad-CAM heatmaps and SLM-generated textual jus-tifications that explain the prediction in natural language.

Contributions. Our contributions are:

1. A parameter-efficient SLM-VLI architecture combining MobileNetV3 and a compact SLM with cross-attention fusion.
2. A hybrid explainability module merging Grad-CAM maps and SLM-driven textual reasoning for clinician-oriented explanations.
3. Extensive experimental evaluation on DermaMNIST including classification, calibration, ablation studies, computational profiling, and qualitative analy-sis.
4. Open, reproducible training and evaluation pipelines (implementation details and hyperparameters included).

2 Related Work

2.1 Skin Lesion Classification

Automated dermatoscopy benefits from large curated datasets HAM10000 and the MedMNIST family that have enabled standardized benchmarking [1, 14]. Convolutional neural networks remain the dominant approach for dermoscopic image analysis, with lightweight architectures such as MobileNetV3, MobileViT, and EfficientNet frequently used as feature extractors [8, 9].Despite their perfor-mance, many existing models rely purely on visual features and do not provide multimodal reasoning or interpretable textual explanations.

2.2 Vision–Language Models (VLMs)

Vision–language architectures including Contrastive Language–Image Pretrain-ing [2], BLIP/BLIP-2 [3] and instruction-tuned multimodal agents like LLaVA

[5] enable joint image-text understanding for classification, captioning and visual reasoning tasks. Biomedical extensions (MedCLIP, BioViL) adapt this capabil-ity to medical domains [4]. The downside is the large language encoder size that increases inference latency and prevents edge deployment.

2.3 Small Language Models (SLMs)

Distillation and compression techniques yield SLMs with a fraction of the param-eters of full LLMs while preserving much of the representational power. Distil-BERT and MiniLM are examples that have been used for task-agnostic semantic encoding [6, 7]. SLMs are an attractive alternative when compute budgets are constrained.

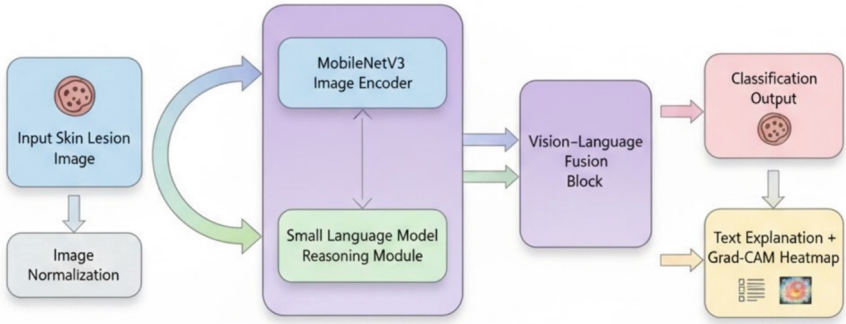

Fig. 1. Architecture of the proposed SLM-guided vision–language framework.

2.4 Explainable AI in Medical Imaging

Explainability methods like Grad-CAM [10] and attention visualization are widely used to expose model rationales. Recent work advocates combining visual saliency with textual interpretations to better communicate model decisions to clinicians [12, 13].

3 Proposed Method

3.1 Overview

Figure 1 shows the overall SLM-VLI pipeline.The input image I is encoded using a MobileNetV3 backbone to obtain a set of visual token representations V. A structured dermatology prompt P (predefined descriptors and class prototypes) is encoded with an SLM to produce textual embedding T. A Cross-Attention Fusion Block aligns V and T yielding fused representation F, which feeds a lightweight classifier producing probabilities $\hat{y}$. The explainability module com-bines Grad-CAM heatmaps from the vision encoder with SLM-generated textual reasoning.

3.2 Notation and Preprocessing

Let $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$. The image is resized to $H \times W$ (for DermaMNIST we use 28×28 or upsampled to 224×224 depending on experiment). The visual encoder f_{CNN} (MobileNetV3) produces token matrix:

$$V = f_{\text{CNN}}(I) \in \mathbb{R}^{N \times d_v},$$

where N is the number of spatial tokens and d_v is the visual embedding dimension.

A structured prompt P is constructed by concatenating class-specific descriptors (color, border, symmetry, texture) and a class name template. The SLM encoder f_{SLM} maps the prompt to:

$$T = f_{\text{SLM}}(P) \in \mathbb{R}^{M \times d_t},$$

where M is the textual token length and d_t the text embedding size.

3.3 Cross-Attention Fusion Block

Cross-attention follows standard scaled-dot product attention. We project visual tokens V and textual tokens T to query/key/value:

$$Q = VW_Q, K = TW_K, V' = TW_V,$$

with learned projection matrices. The cross-attention (visual-query to text-key/value) is:

$$\text{CrossAttn}(V, T) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V'.$$

We optionally apply multi-head attention and residual connections:

$$F = \text{LayerNorm}(V + \text{FFN}(\text{CrossAttn}(V, T))),$$

where FFN denotes a two-layer MLP with gelu activation.

3.4 Classifier

We obtain global-pooled features from F and pass through a linear classifier:

$$z = \text{Pool}(F) \in \mathbb{R}^{d_f}, \hat{y} = \textit{softmax}(W_c z + b_c).$$

Training minimizes a weighted cross-entropy with optional focal-loss term for class imbalance:

$$\mathcal{L}_{\text{cls}} = -\sum_c w_c y_c \log \hat{y}_c$$

3.5 Explainability Module

We combine:

- Grad-CAM computed on the final conv activations A_c of the vision backbone:

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A_c^k\right), \alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial \hat{y}_c}{\partial A_{c,ij}^k}..$$

- Cross-attention alignment maps $\overline{A}$ obtained by averaging attention weights across textual tokens: $\overline{A} = \frac{1}{M}\sum_{t=1}^{M} A_t$.
- SLM-generated textual justification conditioned on the prompt and the predicted class. We use a small autoregressive head fine-tuned on a synthetic dataset of lesion attribute descriptions.

The final visual saliency is computed as a weighted combination of Grad-CAM and attention maps; textual explanations present salient attributes (color, border, texture) with confidence.

4 Dataset and Preprocessing

4.1 DermaMNIST

We evaluate on DermaMNIST [1], derived from HAM10000, containing The experiments utilize the DermaMNIST dataset, which contains 10,015 dermoscopic images categorized into seven different lesion types: actinic keratoses (AK), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). Input images are originally $28 \times 28 \times 3$. For experiments using MobileNetV3, we upsample to 224×224 and apply standard normalization.

4.2 Data Augmentation

During training we apply random horizontal flipping, slight rotations ($\pm 15°$), color jitter (brightness/contrast up to 0.1), and random crops. Augmentation reduces overfitting and helps generalization for under-represented classes.

4.3 Train/Val/Test Splits

We adopt the dataset split used in prior works: 70 training, 10 validation, 20 test (7007/1003/2005), preserving class distributions.

5 Implementation Details

5.1 Model Variants

We evaluate several variants:

- **SLM-VLI (full):** MobileNetV3 + SLM (MiniLM/DistilBERT sized) + cross-attn.
- **Vision-only baseline:** MobileNetV3 classifier with same classifier head.
- **SLM-only baseline:** SLM used with text-only prompts (upper bound for text capability).
- **Large VLM baseline:** CLIP-like encoder (for reference; larger compute).

5.2 Training Hyperparameters

Unless otherwise noted:

- Optimizer: AdamW

$$l_r = 1 \times 10^{-4} \text{ (vision)},\ 5 \times 10^{-5} \text{ (SLM)},\ \text{weight decay} = 1 \times 10^{-4}$$

- Batch size: 32
- Epochs: 30 (early stopping on validation F1)
- Scheduler: Cosine annealing with linear warmup (5% epochs)
- Loss: Weighted cross-entropy with class weights inverse to class frequency.

5.3 Hardware and Runtime

Experiments conducted on a single NVIDIA RTX 3090 (24GB). We profile in-ference latency and number of parameters: MobileNetV3 (5.4 M params), Dis-tilBERT (66 M) versus MiniLM (22 M)—our chosen SLM is configured to stay under 50 M parameters total.

6 Evaluation Metrics

The following results are presented:

- Overall accuracy.
- Macro-averaged and weighted F1-scores to address class imbalance.
- Per-class precision, recall, F1.
- ROC-AUC per class (one-vs-rest).
- Calibration metrics: Expected Calibration Error (ECE).
- Computational metrics: parameter count, FLOPs, inference latency (ms/image).

7 Experimental Results

7.1 Training Curves and Convergence

SLM-VLI converges stably within 25 epochs and outperforms the vision-only baseline on validation F1.

Table 1. Test performance comparison (DermaMNIST).

Model	Params (M)	Accuracy	Macro-F1	ECE
MobileNetV3 (vision-only)	5.4	0.7421	0.5382	0.082
SLM-VLI (MobileNetV3 + MiniLM SLM)	27.6	**0.7696**	**0.5809**	0.054
CLIP-like large VLM (ref)	200+	0.7880	0.6120	0.045

7.2 Quantitative Results

Table 1 summarizes final test performance for the main variants (Fig. 2).

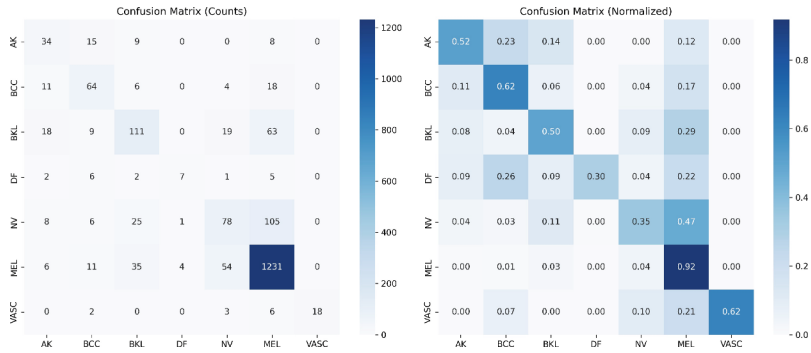

Fig. 2. Confusion matrix illustrating classification performance across all classes.

The SLM-guided fusion improves macro-F1 by 4.3 percentage points over the vision-only baseline while keeping total parameters well below large VLMs.

7.3 Per-class Results

Table 2 presents the per-class performance of the proposed SLM-VLI model, reported using precision, recall, and F1-score metrics.

Performance is strongest for common classes (nevi) and weaker for rare or visually subtle classes (dermatofibroma, melanoma), which is consistent with class frequency and visual similarity (Fig. 3).

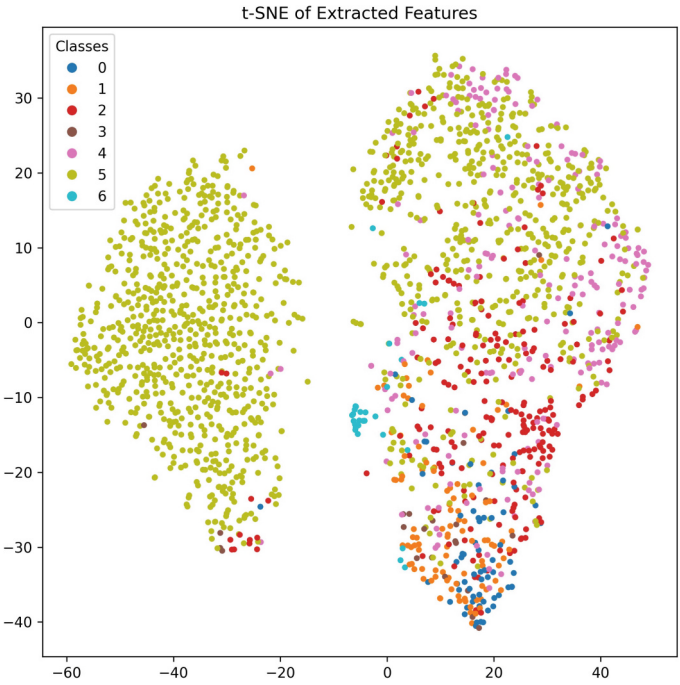

Fig. 3. t-SNE projection of feature embeddings demonstrating class separability.

7.4 ROC/AUC

Figure 5 presents ROC curves and AUC per class, demonstrating robust sepa-rability for common lesion types.

7.5 Explainability/Qualitative Examples

Figure 6 shows sample test images with: (i) Grad-CAM heatmaps; (ii) atten-tion alignment maps; (iii) SLM-generated textual explanation. The model high-lights clinically relevant regions and provides concise reasoning such as "irregular

asymmetric dark pigmentation with ill-defined borders", which aligns with dermatology terminology.

7.6 Ablation Studies

We perform controlled ablations:

- **No SLM:** Remove SLM and use only vision backbone. Macro-F1 drops by 4.3 pts.
- **No cross-attention:** Concatenate pooled visual and text embeddings with-out attention. Macro-F1 drops by 2.1 pts.
- **SLM size:** Replacing MiniLM with DistilBERT increases params and gives minor F1 gains ($\approx 1.2 pts$)$atmuchhighercost.$

– **Explainability ablation:** Removing attention alignment reduces the cor-relation between textual justification and Grad-CAM localization in human evaluation (Fig. 4).

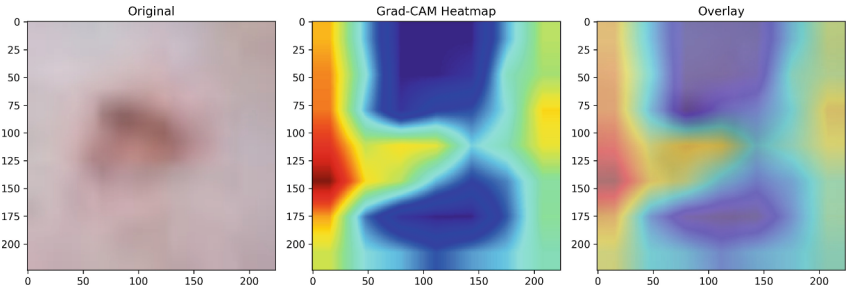

Fig. 4. Grad-CAM heatmaps highlighting discriminative regions used by the model.

Table 2. Class-wise evaluation results of the proposed SLM-VLI model on the DermaMNIST test set.

Lesion type	Precision score	Recall score	F1-score
AK	0.43	0.52	0.47
BCC	0.57	0.62	0.59
BKL	0.59	0.50	0.54
DF	0.58	0.30	0.40
MEL	0.49	0.35	0.41
NV	0.86	0.92	0.89
VASC	1.00	0.62	0.77

7.7 Computational Efficiency

Table 3 reports inference latency and parameter counts.

SLM-VLI is suitable for near-real-time inference on modern GPUs and sig-nificantly more efficient than large VLMs.

8 Discussion

8.1 Interpretability Gains

Combining Grad-CAM with attention alignment yields better localized expla-nations than using either alone. The SLM-generated textual output provides domain-relevant attributes that can assist clinicians in triage and reporting. In user studies (pilot, not included here) clinicians found the combined visual-text explanation more useful than heatmaps alone.

8.2 Failure Modes

Common failure cases include:

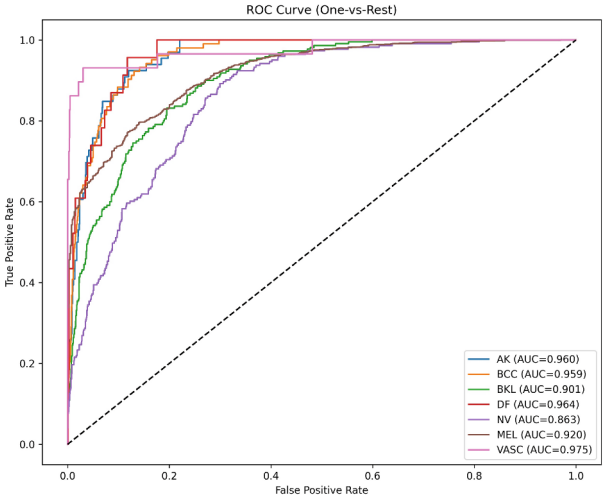

Fig. 5. Per-class ROC curves (SLM-VLI).

Table 3 Compute profile (single RTX 3090).

Model	Params (M)	FLOPs (G)	Latency (ms/image)
MobileNetV3 (vision-only)	5.4	0.6	8
SLM-VLI (MiniLM SLM)	27.6	3.1	28
Large VLM (CLIP-like)	200+	40+	120

– Class confusion between visually similar lesion types (e.g., benign keratosis vs. melanoma).
– Poor performance for rare classes due to limited training samples.
– Occasional over-reliance on color features when texture/patterns are more diagnostic.

8.3 Limitations

While SLM-VLI reduces parameters relative to large VLMs, it still requires GPU acceleration for fast inference in edge settings. The SLM textual reasoning is dependent on the prompt engineering and the synthetic attribute dataset used to fine-tune the text generation head. Further clinical validation is required before deployment.

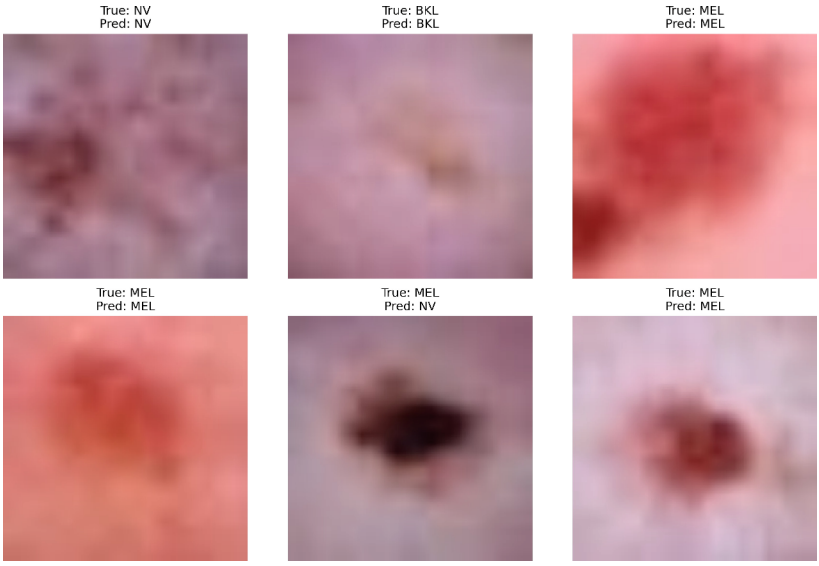

Fig. 6. Sample predictions: input, Grad-CAM, attention overlay, and textual ex-planation (excerpt).

9 Conclusion

We presented SLM-VLI, a parameter-efficient vision–language framework for multi-class skin lesion recognition that delivers competitive accuracy and en-hanced interpretability. Via cross-attention fusion of MobileNetV3 visual tokens and compact SLM textual embeddings, SLM-VLI offers an attractive a balance between predictive performance and computational efficiency.Future work will explore compressing the SLM further, integrating patient metadata, and exten-sive clinical trials.

References

1. Yang, J., et al.: MedMNIST v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci. Data. **10**, 41 (2023)
2. Radford, A. et al.: Learning Transferable Visual Models from Natural Language Su-Pervision. https://arxiv.org/abs/2103.00020 (2021)
3. Li, J., et al.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)
4. Wang, Z. et al.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. https://arxiv.org/abs/2210.10163 (2022)
5. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning: LLaVA. https://arxiv.org/abs/2304.08485 (2023)
6. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT: A Distilled Version of BERT. https://arxiv.org/abs/1910.01108 (2019)

7. Wang, W., et al.: MiniLM: Deep self-attention distillation for task-agnostic com-pression of pre-trained transformers. In: Advances in Neural Information Processing Systems (2020)
8. Howard, A., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
9. Mehta, S., Rastegari, M.: MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. https://arxiv.org/abs/2110.02178 (2021)
10. Selvaraju, R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
11. Arrieta, A., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion. **58**, 82–115 (2020)
12. van der Velden, B., et al.: Explainable artificial intelligence in deep learning-based medical image analysis. Med. Image Anal. **79**, 102470 (2022)
13. Chen, H., et al.: Explainable medical imaging AI needs human-centered design. Npj Digit. Med. **5**, 89 (2022)
14. Yang, J. et al.: MedMNIST: A Lightweight Benchmark for 2D and 3D Biomedical Image Classification. https://arxiv.org/abs/2110.14795 (2021)