Retrieval-Augmented VLMs for Multimodal Melanoma Diagnosis

Jihyun Moon[0009−0009−4373−2600] and Charmgil Hong[0000−0002−8176−252X]

Handong Global University, Pohang, Republic of Korea
{jhmoon, charmgil}@handong.ac.kr

Abstract. Accurate and early diagnosis of malignant melanoma is critical for improving patient outcomes. While convolutional neural networks (CNNs) have shown promise in dermoscopic image analysis, they often neglect clinical metadata and require extensive preprocessing. Vision-language models (VLMs) offer a multimodal alternative but struggle to capture clinical specificity when trained on general-domain data. To address this, we propose a retrieval-augmented VLM framework that incorporates semantically similar patient cases into the diagnostic prompt. Our method enables informed predictions without fine-tuning and significantly improves classification accuracy and error correction over conventional baselines. These results demonstrate that retrieval-augmented prompting provides a robust strategy for clinical decision support.

Keywords: Vision-Language Model · Retrieval-Augmented Generation · Melanoma Diagnosis · Classification Task.

1 Introduction

Malignant melanoma is the most common and deadliest form of skin cancer, with 100,640 new cases and 8,290 deaths reported in the United States in 2024 [16]. Early detection significantly improves survival, which emphasizes the need for accurate and timely diagnosis. Automated diagnostic tools can assist clinicians in detecting malignant lesions at an earlier stage. This can lead to improved prognosis and make timely intervention more achievable. While convolutional neural network (CNN)-based methods have shown promise [10, 11], most rely solely on dermoscopic images and often require preprocessing steps such as region of interest (ROI) segmentation, which limits their utility in clinical practice.

To address these limitations, recent efforts have explored multimodal frameworks that incorporate both images and clinical metadata have gained attention for improving diagnostic accuracy and personalization. Vision-language models (VLMs) [13, 2] have emerged as strong candidates for such tasks, as they jointly process visual and textual data without the need for handcrafted preprocessing. However, off-the-shelf VLMs that are trained on general-purpose data often fail to capture domain-specific complexities [5]. Although fine-tuning with clinical data can mitigate this issue, it requires curated datasets and significant computational resources. As a result, this approach is often infeasible due to privacy constraints and institutional variability.

As an alternative, retrieval-augmented generation (RAG) [12] provides external knowledge-based inference by retrieving similar patient cases and incorporating them into prompts. This approach is particularly appealing for clinical applications, since it enables reasoning without modifying model weights. Previous studies in content-based image retrieval (CBIR) have shown that case-based reasoning can support dermatological diagnosis by referencing visually similar lesions [19, 3]. Building on this idea, our work extends case-based reasoning to a multimodal context by integrating retrieved image–text pairs into VLM prompts. This design supports clinical reasoning by reflecting the way how physicians interpret new cases through analogical comparison with prior patient examples.

More specifically, this study proposes a multimodal diagnostic framework that integrates RAG into a VLM to support more accurate and clinically relevant melanoma classification. We investigate whether retrieved examples improve diagnostic decisions, particularly in correcting false positives and false negatives. Through comprehensive experiments, we show that our method consistently outperforms conventional classification models in both accuracy and error correction. Our main contributions are as follows:

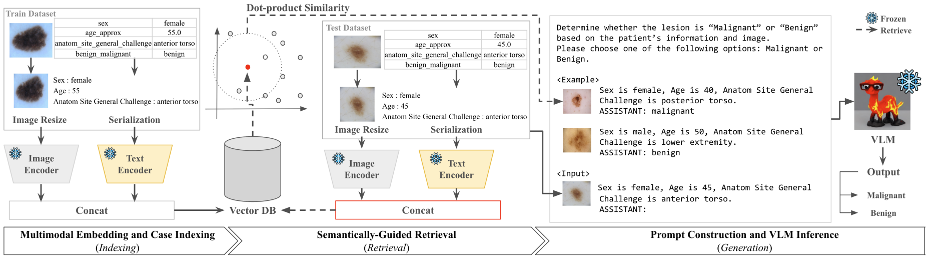
- We propose a retrieval-augmented VLM-based diagnostic framework for melanoma classification by incorporating image–metadata–label examples into prompts to improve decision accuracy.
- We evaluate the effects of different metadata serialization strategies and image encoders on retrieval effectiveness and diagnostic performance.
- We show that the proposed method consistently outperforms conventional image-based, text-based, and early-fusion baselines across multiple metrics and architectures, without requiring fine-tuning.

2 Proposed Approach

This section presents a multimodal diagnostic framework that combines a VLM with RAG to classify melanoma using both dermoscopic images and clinical metadata. In clinical practice, diagnosis often involves comparing a case with prior cases that share similar visual features or clinical attributes. This case-based reasoning improves diagnostic accuracy by using past experience.

Our framework incorporates a retrieval module that searches a database of dermoscopic images and metadata to find semantically similar cases. These examples serve as clinical references and provide contextual support for the prediction of the model. By embedding the retrieved cases into the VLM input prompt, the system emulates the comparative reasoning process used in human diagnosis. This design addresses the limitations of general-purpose VLMs and better aligns the model with the specific demands of melanoma classification. An overview of the architecture is shown in Fig. 1.

Multimodal Embedding and Case Indexing Each training sample includes a dermoscopic image and associated metadata (*e.g.*, age, sex, and lesion location). To process these modalities, we use modality-specific encoders: CNN-based backbones (*e.g.*, ResNeXt-50 [22], EfficientNet-V2-M [17]) for images and


Train Dataset
sex | female
age_approx | 55.0
anatom_site_general_challenge | anterior torso
benign_malignant | benign
Sex : female
Age : 55
Anatom Site General Challenge : anterior torso
Image Resize
Serialization
Image Encoder
Text Encoder
Concat
Dot-product Similarity
Vector DB
Test Dataset
sex | female
age_approx | 45.0
anatom_site_general_challenge | anterior torso
benign_malignant | benign
Sex : female
Age : 45
Anatom Site General Challenge : anterior torso
Image Resize
Serialization
Image Encoder
Text Encoder
Concat
Frozen
Retrieve
VLM
Output
Malignant
Benign
Multimodal Embedding and Case Indexing
(Indexing)
Semantically-Guided Retrieval
(Retrieval)
Prompt Construction and VLM Inference
(Generation)

```
Determine whether the lesion is “Malignant” or “Benign”
based on the patient’s information and image.
Please choose one of the following options: Malignant or
Benign.

<Example>
Sex is female, Age is 40, Anatom Site General
Challenge is posterior torso.
ASSISTANT: malignant

Sex is male, Age is 50, Anatom Site General
Challenge is lower extremity.
ASSISTANT: benign

<Input>
Sex is female, Age is 45, Anatom Site General
Challenge is anterior torso.
ASSISTANT:
```

Fig. 1: Overview of the proposed retrieval-augmented classification framework incorporating attribute-value pair-based prompting.

BERT [8] for text. Metadata is serialized into natural language using template-based sentence transformations (see below) to improve compatibility with language models. The resulting image and text embeddings are concatenated into a single multimodal vector and stored in a FAISS [9] vector database for efficient approximate nearest neighbor search. This setup enables scalable indexing of large dermatological datasets and allows seamless updates as new data is added.

Template-Based Sentence Transformations To make structured metadata compatible with VLMs, we convert patient records into natural language using predefined templates. We apply three serialization strategies:

- **Sentence format**: Expresses each field as a simple sentence (*e.g.*, Age is 45, Sex is female.), matching the VLM's training style.
- **Attribute-value pair**: Uses compact key-value pairs (*e.g.*, Age: 45, Sex: Female) to reduce prompt length and improve parsing.
- **HTML format**: Encodes tabular structure with tags like <table>, <tr>, <th>, and <td> to retain column semantics.

Semantically-Guided Retrieval At inference time, the query sample (image and metadata) is encoded with the same modality-specific encoders used to index the database. We compute dot-product similarity between the query and database embeddings to retrieve the top-K most similar cases. These examples, selected based on both visual and clinical similarity, provide contextual support for VLM prompting. This retrieval introduces domain-specific knowledge without updating model weights and enables adaptation to the target domain. We found that $K = 2$ offers the best balance between contextual relevance and noise.

Prompt Construction and VLM Inference To adapt general-purpose VLMs for binary melanoma classification, we design structured prompts consists of: (1) instruction specifying the task; (2) K retrieved examples as image–metadata–label triplets; and (3) the query sample with a classification request.

The frozen VLM processes the prompt and generates a textual response indicating the predicted class. This few-shot design mirrors clinical reasoning by analogy and leads to better contextual understanding and more reliable predictions. Unlike early-fusion [18] or naive multimodal concatenation [14], our

Table 1: Comparison of image(I)-based, metadata(M)-based, early-fusion, zero-shot VLM, and our proposed framework (bold: the highest performance).

	I	M	Model	Serialization	Accuracy	Balanced Accuracy	Precision	Sensitivity	F1-score	TN	TP	FN	FP
Single Modality	✓	-	ResNeXt-50 [22]	-	0.7380	0.5054	0.2022	0.1316	0.1594	6402	223	1472	880
	✓	-	EfficientNet-V2-M [17]	-	0.6954	0.5061	0.1985	0.2018	0.2001	5901	342	1353	1381
	-	✓	RF [4]	-	0.8156	0.5209	0.6667	0.0472	0.0882	7242	80	1615	40
	-	✓	Vicuna 7B v1.5 [23]	HTML	0.6547	0.4873	0.1725	0.2183	0.1927	5507	370	1325	1775
	-	✓	Vicuna 7B v1.5	Attribute-value pair	0.737	0.5263	0.2294	0.2413	0.2352	5908	409	1286	1374
	-	✓	Vicuna 7B v1.5	Sentence	0.6063	0.5152	0.2023	0.3687	0.2613	4818	625	1070	2464
Early-Fusion with RF	✓	✓	BERT [8]+ResNeXt-50	HTML	0.8607	0.6557	0.8366	0.3263	0.4694	7174	553	1142	108
	✓	✓	BERT+ResNeXt-50	Attribute-value pair	0.8623	0.6568	0.8536	0.3268	0.4277	7187	554	1141	95
	✓	✓	BERT+ResNeXt-50	Sentence	0.8622	0.6589	0.8428	0.3322	0.4765	7177	563	1132	105
	✓	✓	BERT+EfficientNet-V2-M	HTML	0.8501	0.6208	0.8442	0.2525	0.3887	7203	428	1267	79
	✓	✓	BERT+EfficientNet-V2-M	Attribute-value pair	0.8501	0.6220	0.8375	0.2555	0.3915	7198	433	1262	84
	✓	✓	BERT+EfficientNet-V2-M	Sentence	0.8514	0.6232	**0.8546**	0.2566	0.3947	7208	435	1260	74
Early-Fusion with FNN	✓	✓	BERT [8]+ResNeXt-50	HTML	0.6819	0.5079	0.2000	0.2283	0.2132	5734	387	1308	1548
	✓	✓	BERT+ResNeXt-50	Attribute-value pair	0.7040	0.5089	0.2038	0.1953	0.1995	5989	331	1364	1293
	✓	✓	BERT+ResNeXt-50	Sentence	0.7029	0.5009	0.1904	0.1764	0.1832	6011	299	1396	1271
	✓	✓	BERT+EfficientNet-V2-M	HTML	0.7024	0.4967	0.1830	0.1664	0.1743	6023	282	1413	1259
	✓	✓	BERT+EfficientNet-V2-M	Attribute-value pair	0.7084	0.5063	0.2001	0.1817	0.1905	6051	308	1387	1231
	✓	✓	BERT+EfficientNet-V2-M	Sentence	0.7108	0.5090	0.2050	0.1847	0.1943	6068	313	1382	1214
Zero-Shot VLM	✓	✓	LLaVA 7B v1.5 hf [13]	HTML	0.5845	0.6113	0.2608	0.6543	0.3729	4138	1109	586	3144
	✓	✓	LLaVA 7B v1.5 hf	Attribute-value pair	0.7126	0.6128	0.3171	0.4525	0.3729	5630	767	928	1652
	✓	✓	LLaVA 7B v1.5 hf	Sentence	0.5610	0.5658	0.2320	0.5735	0.3303	4064	972	723	3218
Ours ($K = 2$)	✓	✓	BERT+ResNext-50	HTML	0.7396	0.7202	0.3921	**0.6891**	0.4998	5471	1168	527	1811
	✓	✓	BERT+ResNext-50	Attribute-value pair	**0.8876**	**0.7970**	0.7254	0.6513	**0.6864**	6864	1104	591	418
	✓	✓	BERT+ResNext-50	Sentence	0.8810	0.7891	0.7027	0.6413	0.6706	6822	1087	608	460
	✓	✓	BERT+EfficientNet-V2-M	HTML	0.7123	0.6746	0.3505	0.6142	0.4463	5353	1041	654	1929
	✓	✓	BERT+EfficientNet-V2-M	Attribute-value pair	0.8491	0.7345	0.6114	0.5504	0.5793	6689	933	762	593
	✓	✓	BERT+EfficientNet-V2-M	Sentence	0.8459	0.7294	0.6022	0.4322	0.5706	6675	919	776	607

method maintains modality alignment and exploits the ability of the VLM to perform implicit multimodal reasoning.

3 Experiment

Dataset We use the SIIM-ISIC 2019 Challenge dataset [20, 6, 7], which includes 29,923 dermoscopic images with clinical metadata. Among them, 5,608 cases are histopathologically confirmed melanomas. We treat this as a binary classification task: malignant vs. benign. Each sample includes an image and metadata (*age*, *sex*, and *anatomical site*). Images are provided in JPEG format and resized to 224×224 RGB. We apply two-stage stratified sampling to preserve class balance: 70% of the data is used for training and 30% for testing. The training set is further split 80:20 for validation during hyperparameter tuning.

Experimental Setup We evaluate performance under five settings: image-based, text-based, multimodal early-fusion, zero-shot VLM, and our retrieval-augmented VLM framework. For image-based models, we fine-tune ResNeXt-50 [22] and EfficientNet-V2-M [17], both initialized with ImageNet weights. Training uses the Adam optimizer with binary cross-entropy loss, and hyperparameters are tuned via Optuna [1]. For text-based classification, we use Random Forest (RF) [4] and 4-bit quantized Vicuna 7B v1.5 [23], implemented with Scikit-learn [15] and Hugging Face Transformers [21]. In the early-fusion baseline, we extract image features from the final CNN layer and use the *[CLS]* token from the 11th layer of BERT, following the approach in [8]. The two representations are concatenated and classified using either an RF or a ReLU-activated feedforward neural network (FNN). Hyperparameters are tuned via Grid Search [15]

Table 2: Comparison of baseline models and the proposed approach (Ours, $K = 2$), showing the number of corrected errors (FN/FP corrected as TP/TN) and the corresponding recovery rate (%). Recovery is defined as the proportion of corrected errors relative to the total baseline errors (I: Image, M: Metadata).

(a) Performance Comparison Across Different Experimental Settings.

	I	M	Model	Serialization		Ours ($K = 2$)	Recovery (%)
Single Modality	-	✓	RF [4]	-	FP	34	85.00
	-	✓	RF	-	FN	1035	64.09
	-	✓	Vicuna 7B v1.5 [23]	Attribute-value pair	FP	1258	70.87
	-	✓	Vicuna 7B v1.5	Attribute-value pair	FN	827	64.31
Early-Fusion with RF	✓	✓	BERT [8]+ResNeXt-50 [22]	Attribute-value pair	FP	71	74.74
	✓	✓	BERT+ResNeXt-50	Attribute-value pair	FN	604	52.94
Zero-Shot VLM	✓	✓	LLaVA 7B v1.5 hf [13]	Attribute-value pair	FP	1507	91.22
	✓	✓	LLaVA 7B v1.5 hf	Attribute-value pair	FN	571	61.53

(b) Performance Comparison Across Different Serialization Methods.

	I	M	Model	Serialization		Attribute-value pair	Recovery (%)
Ours ($K = 2$)	✓	✓	BERT [8]+ResNeXt-50 [22]	HTML	FP	1513	83.55
	✓	✓	BERT+ResNeXt-50	HTML	FN	116	22.01
Ours ($K = 2$)	✓	✓	BERT+ResNeXt-50	Sentence	FP	113	24.57
	✓	✓	BERT+ResNeXt-50	Sentence	FN	43	7.07

(c) Performance Comparison Across Different Image Encoder.

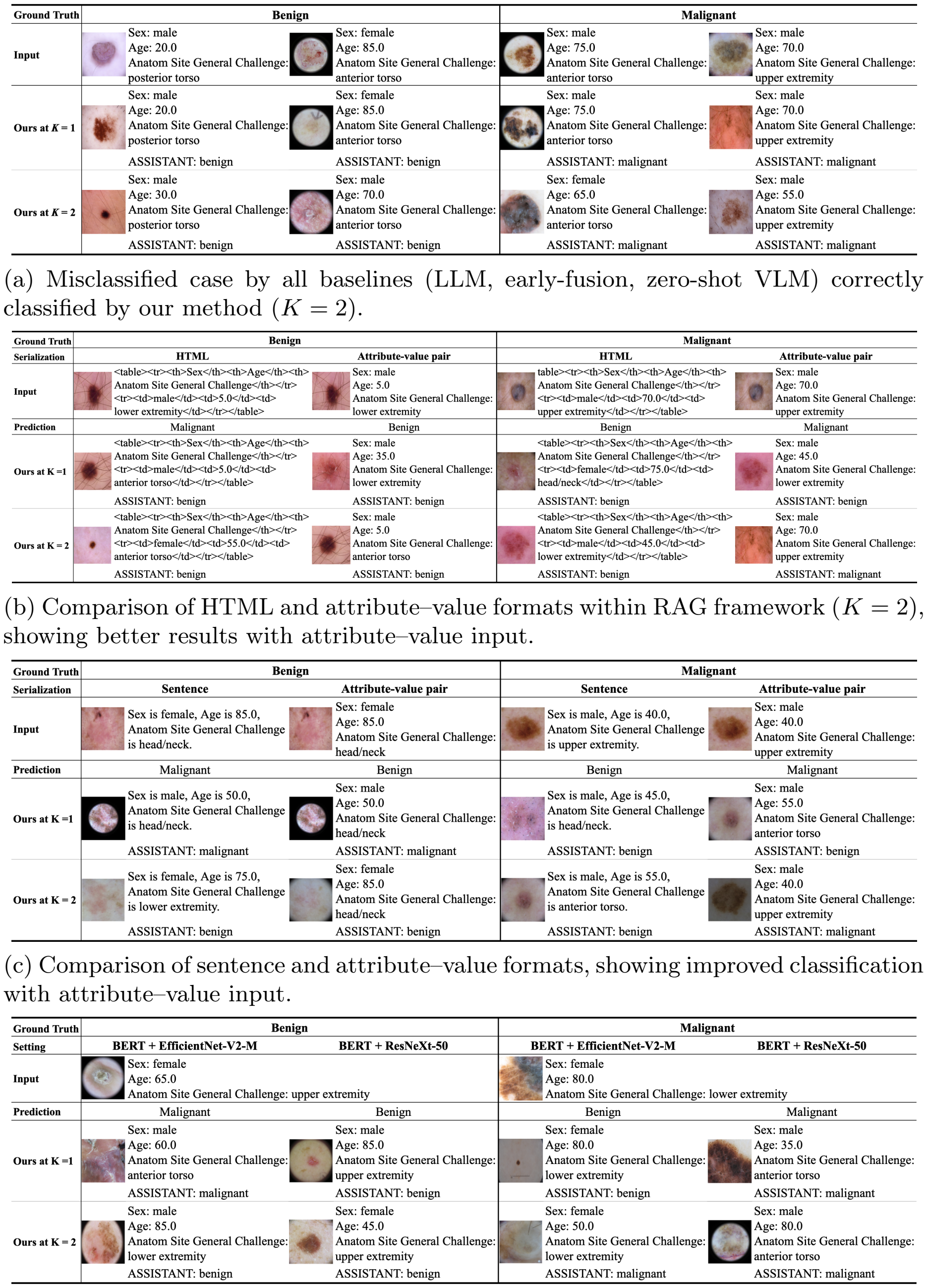
	I	M	Model	Serialization		BERT [8]+ResNeXt-50 [22]	Recovery (%)
Ours ($K = 2$)	✓	✓	BERT+EfficientNet-V2-M [17]	Attribute-value pair	FP	487	82.12
	✓	✓	BERT+EfficientNet-V2-M	Attribute-value pair	FN	325	42.65

or Optuna [1]. For VLM-based experiments, we use the 4-bit quantized version of LLaVA v1.5 [13]. In the RAG configuration, we construct a FAISS [9] index containing 16,756 image–text pairs from the training set. For each query, the top two nearest neighbors ($K = 2$) are retrieved and inserted into the model prompt. We empirically evaluate different values of K ($K = 1, 2, 3, 4$) and find that $K = 2$ offers the best trade-off between contextual relevance and noise.

3.1 Quantitative Evaluation

Table 1 summarizes the classification results. We report accuracy, balanced accuracy, precision, sensitivity, F1-score, and confusion matrix components. Given the class imbalance and the clinical importance of minimizing both false negative (FN) and false positive (FP), we adopt F1-score as the primary metric.

Single-Modality Models Models using only dermoscopic images or clinical metadata show limited diagnostic performance. Among image-based methods, EfficientNet-V2-M achieves the best results, though its performance is affected by visual noise due to the absence of preprocessing. In text-based models, Vicuna



Ground Truth	Benign		Malignant	
Input	Sex: male Age: 20.0 Anatom Site General Challenge: posterior torso	Sex: female Age: 85.0 Anatom Site General Challenge: anterior torso	Sex: male Age: 75.0 Anatom Site General Challenge: anterior torso	Sex: male Age: 70.0 Anatom Site General Challenge: upper extremity
Ours at $K = 1$	Sex: male Age: 20.0 Anatom Site General Challenge: posterior torso ASSISTANT: benign	Sex: female Age: 85.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign	Sex: male Age: 75.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant	Sex: male Age: 70.0 Anatom Site General Challenge: upper extremity ASSISTANT: malignant
Ours at $K = 2$	Sex: male Age: 30.0 Anatom Site General Challenge: posterior torso ASSISTANT: benign	Sex: male Age: 70.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign	Sex: female Age: 65.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant	Sex: male Age: 55.0 Anatom Site General Challenge: upper extremity ASSISTANT: malignant

(a) Misclassified case by all baselines (LLM, early-fusion, zero-shot VLM) correctly classified by our method ($K = 2$).

Ground Truth	Benign		Malignant	
Serialization	HTML	Attribute-value pair	HTML	Attribute-value pair
Input	<table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>5.0</td><td> lower extremity</td></tr></table>	Sex: male Age: 5.0 Anatom Site General Challenge: lower extremity	table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>70.0</td><td> upper extremity</td></tr></table>	Sex: male Age: 70.0 Anatom Site General Challenge: upper extremity
Prediction	Malignant	Benign	Benign	Malignant
Ours at K =1	<table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>5.0</td><td> anterior torso</td></tr></table> ASSISTANT: benign	Sex: male Age: 35.0 Anatom Site General Challenge: lower extremity ASSISTANT: benign	<table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>female</td><td>75.0</td><td> head/neck</td></tr></table> ASSISTANT: benign	Sex: male Age: 45.0 Anatom Site General Challenge: lower extremity ASSISTANT: benign
Ours at K = 2	<table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>female</td><td>55.0</td><td> anterior torso</td></tr></table> ASSISTANT: benign	Sex: male Age: 5.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign	<table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>45.0</td><td> lower extremity</td></tr></table> ASSISTANT: benign	Sex: male Age: 70.0 Anatom Site General Challenge: upper extremity ASSISTANT: malignant

(b) Comparison of HTML and attribute–value formats within RAG framework ($K = 2$), showing better results with attribute–value input.

Ground Truth	Benign		Malignant	
Serialization	Sentence	Attribute-value pair	Sentence	Attribute-value pair
Input	Sex is female, Age is 85.0, Anatom Site General Challenge is head/neck.	Sex: female Age: 85.0 Anatom Site General Challenge: head/neck	Sex is male, Age is 40.0, Anatom Site General Challenge is upper extremity.	Sex: male Age: 40.0 Anatom Site General Challenge: upper extremity
Prediction	Malignant	Benign	Benign	Malignant
Ours at K =1	Sex is male, Age is 50.0, Anatom Site General Challenge is head/neck. ASSISTANT: malignant	Sex: male Age: 50.0 Anatom Site General Challenge: head/neck ASSISTANT: malignant	Sex is male, Age is 45.0, Anatom Site General Challenge is head/neck. ASSISTANT: benign	Sex: male Age: 55.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign
Ours at K = 2	Sex is female, Age is 75.0, Anatom Site General Challenge is lower extremity. ASSISTANT: benign	Sex: female Age: 85.0 Anatom Site General Challenge: head/neck ASSISTANT: benign	Sex is male, Age is 55.0, Anatom Site General Challenge is anterior torso. ASSISTANT: benign	Sex: male Age: 40.0 Anatom Site General Challenge: upper extremity ASSISTANT: malignant

(c) Comparison of sentence and attribute–value formats, showing improved classification with attribute–value input.

Ground Truth	Benign		Malignant	
Setting	BERT + EfficientNet-V2-M	BERT + ResNeXt-50	BERT + EfficientNet-V2-M	BERT + ResNeXt-50
Input	Sex: female Age: 65.0 Anatom Site General Challenge: upper extremity		Sex: female Age: 80.0 Anatom Site General Challenge: lower extremity	
Prediction	Malignant	Benign	Benign	Malignant
Ours at K =1	Sex: male Age: 60.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant	Sex: male Age: 85.0 Anatom Site General Challenge: upper extremity ASSISTANT: benign	Sex: female Age: 80.0 Anatom Site General Challenge: lower extremity ASSISTANT: benign	Sex: male Age: 35.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant
Ours at K = 2	Sex: male Age: 85.0 Anatom Site General Challenge: lower extremity ASSISTANT: benign	Sex: female Age: 45.0 Anatom Site General Challenge: upper extremity ASSISTANT: benign	Sex: female Age: 50.0 Anatom Site General Challenge: lower extremity ASSISTANT: malignant	Sex: male Age: 80.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant

(d) Impact of image encoder choice (EfficientNet-V2-M vs. ResNeXt-50) using attribute–value format in RAG framework.

Fig. 2: Error cases corrected by our framework with retrieved cases.

7B v1.5 outperforms RF, likely benefiting from general-domain pretraining. The RF model struggles to capture patterns effectively due to the limited number of metadata features. These results suggest that single-modality approaches are inadequate for accurate melanoma diagnosis and highlight the importance of multimodal integration.

Multimodal Fusion and Zero-Shot VLM Zero-shot VLM outperforms early-fusion with FNN due to its use of attention-based mechanisms that capture cross-modal interactions. In contrast, FNN relies on simple feature concatenation, which limits its ability to represent semantic relationships. Interestingly, early-fusion with RF achieves better results than zero-shot VLM, indicating that pretrained VLMs do not fully capture clinical signals. These observations point to the need for strategies that incorporate domain knowledge and task-specific examples to improve VLM-based classification.

Proposed RAG Framework Our RAG-based VLM framework achieves the highest performance across all settings. Using BERT, ResNeXt-50, and attribute-value pair serialization, it reaches an F1-score of 0.6864, improving by 0.2099 over the best early-fusion model and by 0.3135 over zero-shot VLM. Sensitivity increases to 0.6513, more than doubling that of early-fusion, while precision reaches 0.7254. These results demonstrate that retrieval-augmented prompting provides consistent gains in both accuracy and clinical error correction, while maintaining a strong balance between precision and recall.

3.2 Qualitative Evaluation

The quantitative results show that the RAG-based VLM framework outperforms image-based, text-based, early-fusion, and zero-shot VLM models. To better understand this performance, we qualitatively examine how retrieved examples contribute to correcting FP and FN that baseline methods fail to resolve.

Error Analysis of Baseline Predictions Each baseline model uses the best-performing architecture for its modality. Our framework ($K = 2$) applies BERT with ResNeXt-50 and attribute–value pair serialization. Table 2a presents recovery rates, defined as the proportion of FP and FN errors that our method correctly reclassifies. Zero-shot VLM achieves recovery rates of 91.22% for FP and 61.53% for FN. In contrast, early-fusion with RF shows substantially lower recovery, likely due to limited capacity to model semantic interactions across modalities. Fig. 2a shows representative cases. In the left column, baseline models misclassify benign lesions as malignant. In the right, malignant lesions are predicted as benign. Our method retrieves clinically similar examples based on *sex*, *age*, and *anatomical site*, and inserts them into the prompt. For instance, two retrieved benign cases from the same *anatomical site* (posterior torso) help correct a prior false positive. This context influences the model's decision and leads to more accurate predictions.

Overall, these results suggest that retrieval-based prompting addresses key limitations of conventional models and supports more reliable clinical reasoning.

Effect of Input Serialization Format To evaluate the impact of input serialization, we compare three formats: HTML, attribute–value pair, and sentence, using the same configuration (BERT + ResNeXt-50, $K = 2$). Table 2b shows recovery rates where FP and FN errors under HTML and sentence formats are corrected to true positive (TP) and true negative (TN) by switching to attribute–value format.

For HTML input, 83.55% of FP and 22.01% of FN errors are corrected. In contrast, conversion from sentence format yields 24.57% for FP and 7.07% for FN. These results suggest that attribute–value input encodes clinical variables more explicitly. Although HTML preserves the same content, its tag-based structure may obscure important features during embedding. While sentence format aligns with VLM training data, its classification performance remains lower than attribute–value format. Fig. 2b and 2c show representative examples. In Fig. 2b, the HTML-based model yields a false positive for a benign lesion. The attribute–value format retrieves benign cases with matching *sex* and *age* and enables correct classification. For the malignant example, the HTML input produces false negatives, whereas attribute–value input retrieves similar malignant lesions and supports accurate prediction. Fig. 2c presents a false positive under sentence format that is corrected by attribute–value input, which retrieves benign cases with matching *sex* and *anatomical site*. Its more explicit structure strengthens contextual alignment and improves classification.

In summary, these results show that structured input formats, especially attribute–value pairs, better support retrieval-based reasoning by clarifying the semantic role of each variable.

Effect of Image Encoder Configuration To assess the influence of image encoder on diagnostic performance, we compare ResNeXt-50 and EfficientNet-V2-M, using the same text encoder (BERT), input format (attribute-value pair), and retrieval setting ($K = 2$). Table 2c shows recovery rates where predictions by the EfficientNet-V2-M model are corrected by the ResNeXt-50 model. The recovery rate for FPs reaches 82.12%, and for FNs 42.65%, indicating that ResNeXt-50 is more effective at correcting errors. Fig. 2d presents examples where the two models produce different outcomes for the same inputs. In the benign case, EfficientNet-V2-M misclassifies the lesion as malignant. In contrast, ResNeXt-50 retrieves benign cases with similar *anatomical site* and *sex*, which results in correct classification. These retrieved cases show strong alignment with the query in both spatial and demographic attributes. EfficientNet-V2-M, by comparison, retrieves cases with lower consistency, which may reduce contextual reliability. A similar difference appears in the malignant case. EfficientNet-V2-M predicts FN, whereas ResNeXt-50 retrieves two malignant cases with similar color and lesion spread, resulting in correct classification. These examples provide clearer visual evidence that supports the decision of the model.

Altogether, the results suggest that ResNeXt-50 captures visual features relevant to melanoma classification more effectively and provides stronger contextual alignment in retrieval-based inference.

4 Conclusion

We presented a retrieval-augmented diagnostic framework that integrates VLMs with case-based prompting for melanoma classification. The framework improves diagnostic performance without fine-tuning by retrieving semantically similar examples and inserting them into the input prompt. Quantitative and qualitative results show improved contextual reasoning and fewer classification errors. Comparisons across serialization formats and encoders confirm the benefit of structured input and ResNeXt-50 visual features. These findings support retrieval-augmented prompting as a robust and generalizable strategy for clinical decision support using pretrained multimodal models. Although the framework shows promising results, its reliance on a single VLM may limit generalizability across diverse diagnostic tasks. Future work may expand this approach to multi-class skin lesion classification and other domains that require multimodal reasoning and greater model flexibility.

Acknowledgments. This research was supported (1) by the MSIT (Ministry of Science and ICT), Korea, under the Global Research Support Program in the Digital Field (RS-2024-00431394), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); (2) by the MSIT and NIPA (National IT Industry Promotion Agency), Korea, under the Development of AI Precision Medical Solution project (Dr. Answer 2.0, S0252-21-1001); and (3) under the High Performance Computing Support project.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Akrout, M., Cirone, K.D., Vender, R.: Evaluation of vision llms gtp-4v and llava for the recognition of features characteristic of melanoma. Journal of Cutaneous Medicine and Surgery **28**(1), 98–99 (2024)
3. Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G., Grana, C.: Supporting skin lesion diagnosis with content-based image retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8053–8060. IEEE (2021)
4. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
5. Chen, J., Jiang, Y., Yang, D., Li, M., Wei, J., Qian, Z., Zhang, L.: Can llms' tuning methods work in medical multimodal domain? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 112–122. Springer (2024)
6. 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)

7. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
9. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
10. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
11. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. Journal of Investigative Dermatology **138**(7), 1529–1538 (2018)
12. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
13. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
14. Liu, K., Li, Y., Xu, N., Natarajan, P.: Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730 (2018)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
16. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA: a cancer journal for clinicians **74**(1), 12–49 (2024)
17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
18. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
19. Tschandl, P., Argenziano, G., Razmara, M., Yap, J.: Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. British Journal of Dermatology **181**(1), 155–165 (2019)
20. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. pp. 38–45 (2020)
22. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
23. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)