21st IEEE International Conference on
Advanced Visual and Signal-Based Systems

Multimodal Clinical Decision Support for Melanoma Diagnosis Using Retrieval-Augmented Generation and Vision-Language Models

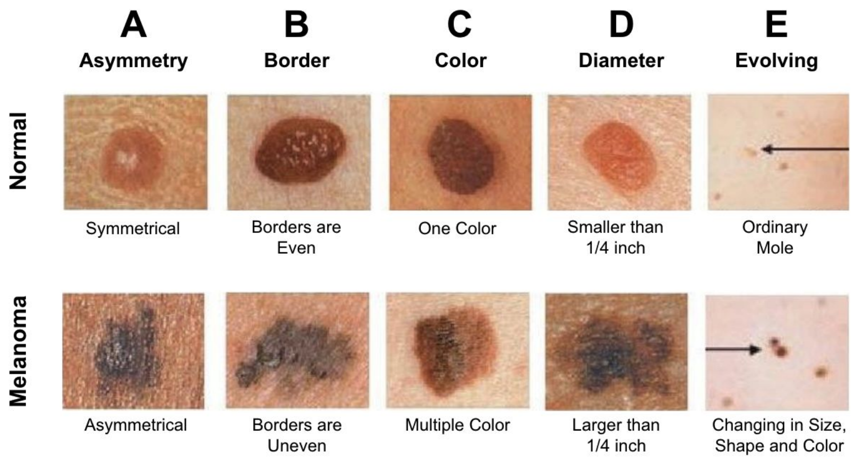
Jihyun Moon , Charmgil Hong

{**jhmoon**, charmgil}@handong.ac.kr

Handong Global University

Introduction

- **Malignant melanoma**
  - **65%** of skin cancer-related mortality
  - Survival rates depend on **early** detection: [Markovic et al., 2007]
    - ➡ Early detection: > 99% 5-year survival rate
    - ➡ After metastasis: **< 35% survival rate**
  - Traditional Diagnosis: The **ABCDE** criteria [Duarte et al., 2021]
    - ➡ **Visual assessment** of lesion's shape, edges, colors, and size

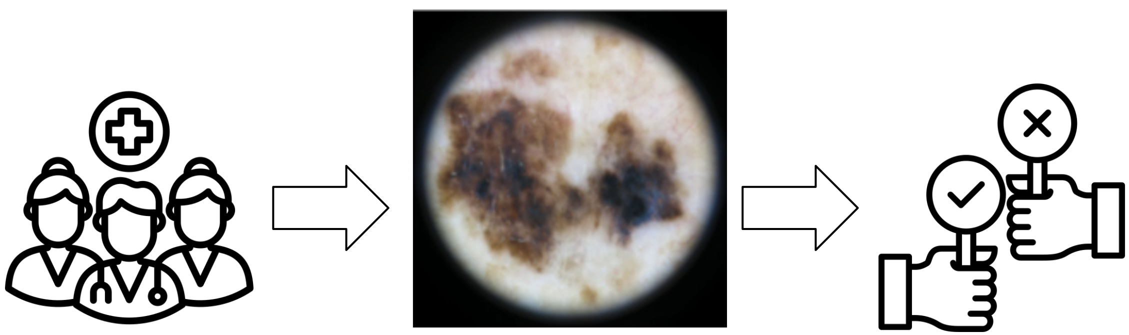


The ABCDEs of Detecting Melanoma [Alafghani, 2018]

Duarte et al. (2021). Clinical abcde rule for early melanoma detection. European Journal of Dermatology.
Markovic et al. (2007). Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis.
Alafghani T. (2018). A CMOS 10-bit SAR ADC, with on-chip offset cancellation, for near-field, mm-wave imaging technique, applied to skin cancer detection.

Introduction

- **Malignant melanoma**
  - Traditional Diagnosis: The ABCDE criteria [Duarte et al., 2021]
    - Problems
      - ➡ **Subjective interpretation** by clinicians
      - ➡ **Experience-dependent** diagnosis
      - ➡ **High variability** between physicians
      - ➡ Potential **inaccuracies** in judgment
    - Need for **automated** and **objective** decision support systems



Raw Dermoscopic image

Duarte et al. (2021). Clinical abcde rule for early melanoma detection. European Journal of Dermatology.
Alafghani T. (2018). A CMOS 10-bit SAR ADC, with on-chip offset cancellation, for near-field, mm-wave imaging technique, applied to skin cancer detection.

Introduction

- From Traditional Assessment to **AI**-based Diagnosis
  - **Convolutional Neural Networks (CNNs)** have advanced automated detection
    - Achieved **dermatologist-level performance** using dermoscopic images [Esteva et al., 2017]
    - Improved robustness and efficiency on **high-resolution inputs** [Han et al., 2018]

➡ Limitations:

- Process only visual data, **ignoring** clinical metadata

- Highly **dependent** on image **processing**

Can **clinical metadata** enhance image-based melanoma diagnosis?

Esteva et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks.
Han et al. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm.
Brinker et al. (2018). Skin cancer classification using convolutional neural networks: systematic review.
Wang et al. (2022). Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images.

Introduction

- From Traditional Assessment to AI-based Diagnosis
  - **Multimodal Fusion** for Melanoma Diagnosis
    - Incorporating demographic information **improves** classification [Brinker et al., 2018]
    - **Attention-based fusion** improves patient-specific prediction [Wang et al., 2022]

➡ Limitations:

- **Weak alignment** between clinical metadata and localized image features

Can a **VLM effectively process** dermoscopic images for diagnostic classification?

Esteva et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks.
Han et al. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm.
Brinker et al. (2018). Skin cancer classification using convolutional neural networks: systematic review.
Wang et al. (2022). Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images.

Introduction

- **Vision-Language Models (VLMs)** in Medical Diagnosis
  - VLMs learn joint embeddings of images and text from large-scale data [Liu et al., 2023]
    - Allow **effective integration** without explicit preprocessing or alignment [Radford et al., 2021]
    - Pretrained on general domain data

Can a VLM achieve **clinically acceptable** diagnostic accuracy?

Liu et al. (2023). Visual instruction tuning.
Radford et al. (2021). Learning transferable visual models from natural language supervision.
Shifai et al. (2024). Can chatgpt vision diagnose melanoma? an exploratory diagnostic accuracy study.
Akrout et al. (2024). Evaluation of vision llms gtp-4v and llava for the recognition of features characteristic of melanoma.

Introduction

- **Vision-Language Models (VLMs)** in Medical Diagnosis
  - VLMs learn joint embeddings of images and text from large-scale data [Liu et al., 2023]
    - Allow **effective integration** without explicit preprocessing or alignment [Radford et al., 2021]
    - Pretrained on general domain data

  ➡ **Lack** sufficient **medical domain** knowledge and clinical context

  - Produced consistent image descriptions but showed **limited** diagnostic accuracy [Akrout et al., 2024]

  - **Inconsistent** sensitivity and specificity raised concerns about clinical reliability [Shifai et al., 2024]

Does **RAG** enhance performance by refining clinical cases **without fine tuning**?

Liu et al. (2023). Visual instruction tuning.
Radford et al. (2021). Learning transferable visual models from natural language supervision.
Shifai et al. (2024). Can chatgpt vision diagnose melanoma? an exploratory diagnostic accuracy study.
Akrout et al. (2024). Evaluation of vision llms gtp-4v and llava for the recognition of features characteristic of melanoma.

Introduction

- Research Questions

Q1. Can **clinical metadata** enhance image-based melanoma diagnosis?

Q2. Can a **VLM effectively process** dermoscopic images for diagnostic classification?

Q3. Can a VLM achieve **clinically acceptable** diagnostic accuracy?

Q4. Does **RAG** enhance performance by refining clinical cases **without fine tuning**?

→ We propose a **multimodal diagnostic framework** that incorporates a **Retrieval-Augmented Generation (RAG)** strategy into a **VLM-based system**.

Liu et al. (2023). Visual instruction tuning.
Radford et al. (2021). Learning transferable visual models from natural language supervision.
Shifai et al. (2024). Can chatgpt vision diagnose melanoma? an exploratory diagnostic accuracy study.
Akrout et al. (2024). Evaluation of vision llms gtp-4v and llava for the recognition of features characteristic of melanoma.

Proposed Framework

- A **retrieval-augmented diagnostic framework** that combines dermoscopic images and clinical metadata for **VLM-based melanoma classification**
  - **Serialization** of tabular metadata
  - Multimodal **indexing** and **retrieval**
  - **Prompt**-based classification with retrieved examples

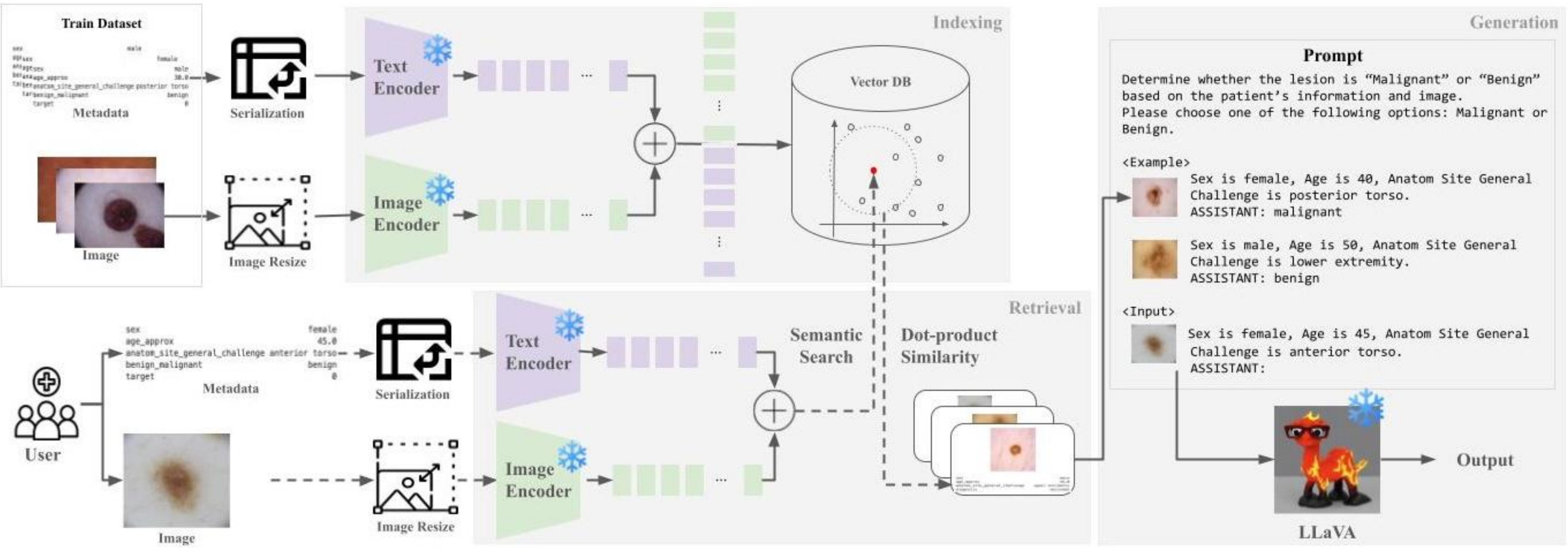


Figure 1. Proposed retrieval-augmented classification framework with sentence-based prompting.

Proposed Framework

- **Prompt-based classification** with **retrieved examples**
  - VLMs are optimized for generative tasks and underperform in discriminative settings

➡ **Design** structured **prompts** that clearly define the task objective and constrain the model output

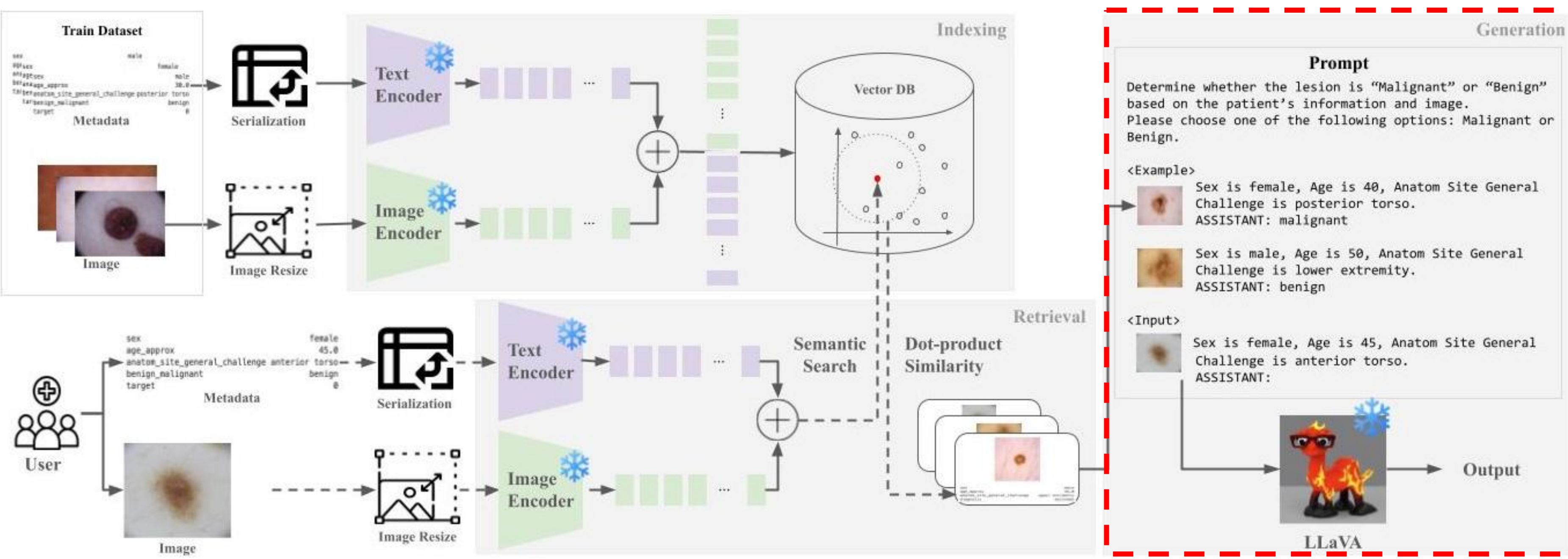


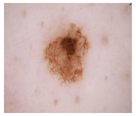
Figure 1. Proposed retrieval-augmented classification framework with sentence-based prompting.

Proposed Framework

- **Prompt****-based classification** with retrieved examples
  - Few-Shot Prompting for Classification
    - **Task Definition**
      - **Clear instruction to classify lesion as “Malignant” or “Benign”**
    - Constrained Output
      - Model must choose between two specific classes
    - Contextual Examples
      - Top-*K* (*K*-shot) retrieved similar cases provide in-context learning
      - Infer the label, resembling the few-shot prompting paradigm
    - Target query
      - Placed under <Input> tag
      - In zero-shot cases, the query is provided without examples

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.

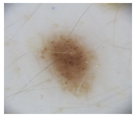
<Example>



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>



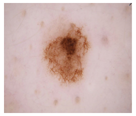
Sex is female, Age is 45, Anatom Site General Challenge is anterior torso.
ASSISTANT:

Proposed Framework

- **Prompt-based classification** with retrieved examples
  - Few-Shot Prompting for Classification
    - Task Definition
      - Clear instruction to classify lesion as “Malignant” or “Benign”
    - **Constrained Output**
      - **Model must choose between two specific classes**
    - Contextual Examples
      - Top-*K* (*K*-shot) retrieved similar cases provide in-context learning
      - Infer the label, resembling the few-shot prompting paradigm
    - Target query
      - Placed under <Input> tag
      - In zero-shot cases, the query is provided without examples

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.

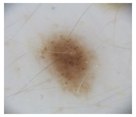
<Example>



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>



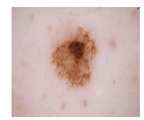
Sex is female, Age is 45, Anatom Site General Challenge is anterior torso.
ASSISTANT:

Proposed Framework

- **Prompt****-based classification** with retrieved examples
  - Few-Shot Prompting for Classification
    - Task Definition
      - Clear instruction to classify lesion as “Malignant” or “Benign”
    - Constrained Output
      - Model must choose between two specific classes
    - **Contextual Examples**
      - **Top-*K* (*K*-shot) retrieved similar cases provide in-context learning**
      - **Infer the label, resembling the few-shot prompting paradigm**
    - Target query
      - Placed under <Input> tag
      - In zero-shot cases, the query is provided without examples

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.

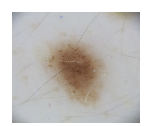
<Example>



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>



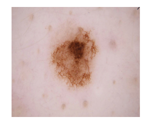
Sex is female, Age is 45, Anatom Site General Challenge is anterior torso.
ASSISTANT:

Proposed Framework

- **Prompt-based classification** with retrieved examples
  - Few-Shot Prompting for Classification
    - Task Definition
      - Clear instruction to classify lesion as “Malignant” or “Benign”
    - Constrained Output
      - Model must choose between two specific classes
    - Contextual Examples
      - Top-*K* (*K*-shot) retrieved similar cases provide in-context learning
      - Infer the label, resembling the few-shot prompting paradigm
    - **Target query**
      - **Placed under <Input> tag**
      - **In zero-shot cases, the query is provided without examples**

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.

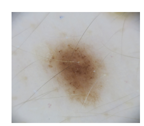
<Example>



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>

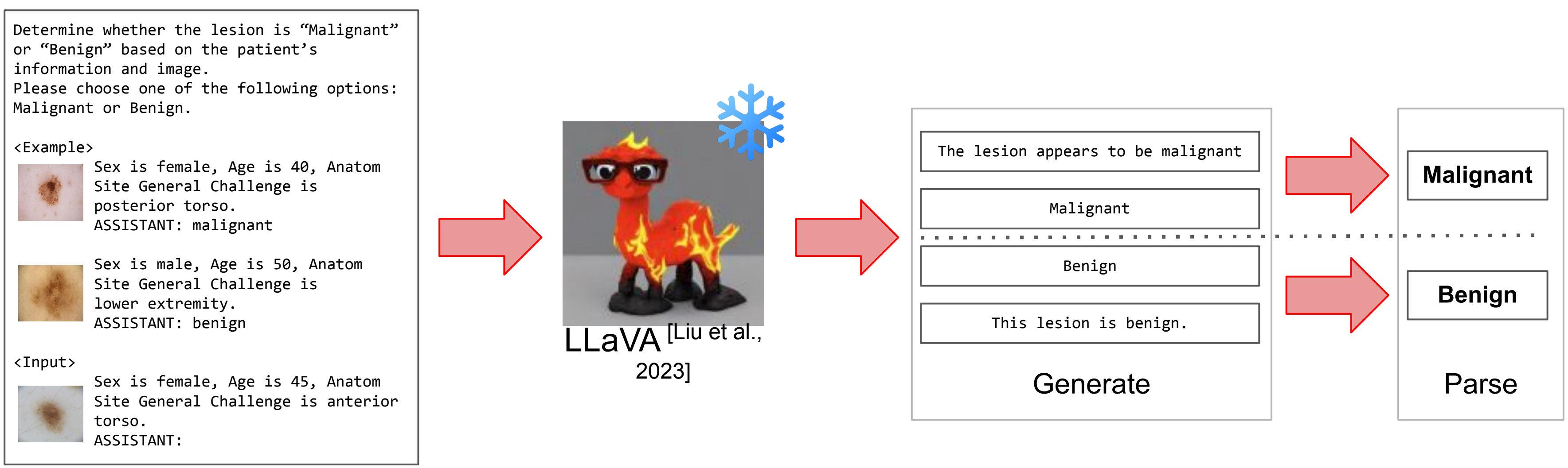


Sex is female, Age is 45, Anatom Site General Challenge is anterior torso.
ASSISTANT:

Proposed Framework

- Prompt-based **classification** with retrieved examples
  - Classification Process
    - **Generate** diagnosis results in **natural language** text form
    - **Parse** to extract sentence containing the keywords “**malignant**” or “**benign**”
    - **Determine** the final classification label

➡ Enable the model to provide natural language explanations while producing label



Liu et al. (2023). Visual instruction tuning.

Proposed Framework

- **Serialization** of tabular metadata
  - Pre-trained VLMs process text-based inputs

➡ **Converting** structured metadata into **natural language** to enable prompting and embedding

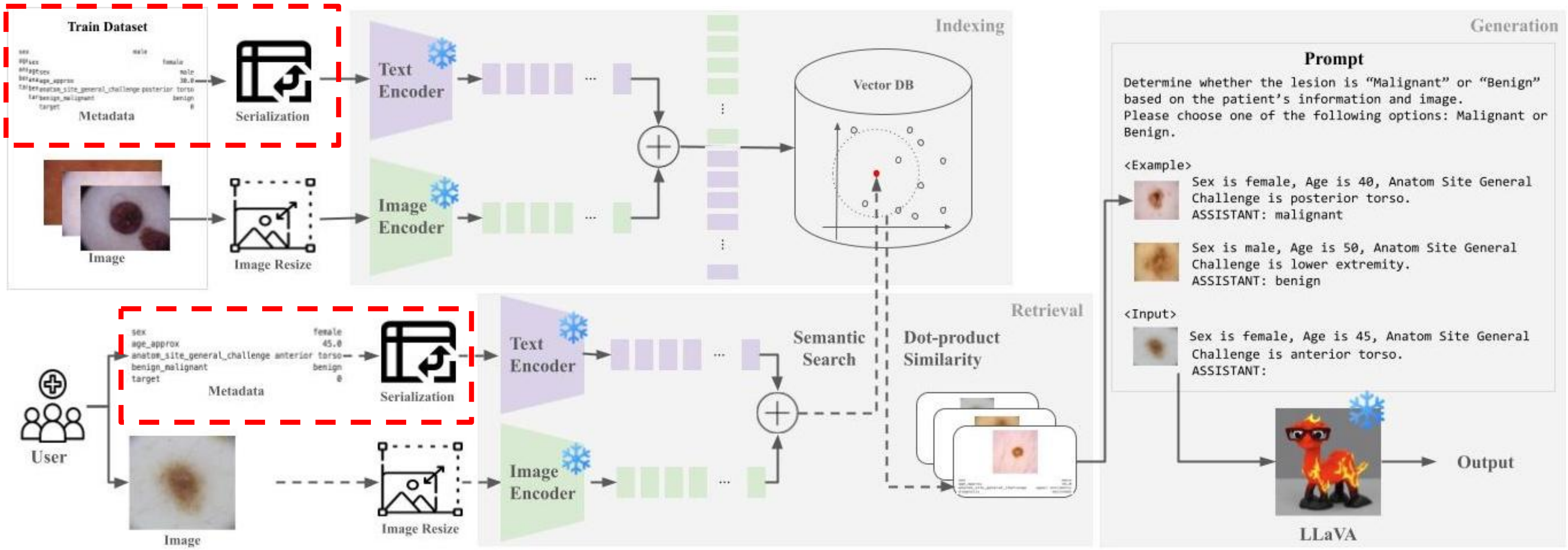


Figure 1. Proposed retrieval-augmented classification framework with sentence-based prompting.

Proposed Framework

- Serialization of Tabular Metadata
  - **Converting** structured clinical metadata into **natural language** for VLM processing
  - 4 serialization approaches explored:
    1. **HTML**: Uses **<table>**, **<th>**, and **<td> tags** to explicitly preserve tabular structure

Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Raw Clinical Metadata

HTML

```
<table>
 <tr>
  <th>Sex</th>
  <th>Age</th>
  <th>Anatomic Site General Challenge</th>
 </tr>
 <tr>
  <td>female</td>
  <td>55</td>
  <td>anterior torso</td>
 </tr>
</table>
```

Proposed Framework

- Serialization of Tabular Metadata
  - **Converting** structured clinical metadata into **natural language** for VLM processing
  - 4 serialization approaches explored:
    1. HTML: Uses <table>, <th>, and <td> tags to explicitly preserve tabular structure
    2. **Markdown**: Formats data as a simple table using **|** and **---** for columns and rows

Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Raw Clinical Metadata

Markdown

```
|  Sex   | Age | Anatom Site General Challenge |
|----------|--------|------------------------------------------|
| female |  55   |          anterior torso                  |
```

Proposed Framework

- Serialization of Tabular Metadata
  - **Converting** structured clinical metadata into **natural language** for VLM processing
  - 4 serialization approaches explored:
    1. HTML: Uses <table>, <th>, and <td> tags to explicitly preserve tabular structure
    2. Markdown: Formats data as a simple table using | and --- for columns and rows
    3. **Attribute-Value pair**: Lists each attribute and its value as a compact **key–value** pair

Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Raw Clinical Metadata

Attribute-Value pair
Sex: female, Age: 55, Anatomic Site General Challenge: anterior torso

Proposed Framework

- Serialization of Tabular Metadata
  - **Converting** structured clinical metadata into **natural language** for VLM processing
  - 4 serialization approaches explored:
    1. HTML: Uses <table>, <th>, and <td> tags to explicitly preserve tabular structure
    2. Markdown: Formats data as a simple table using | and --- for columns and rows
    3. Attribute-Value pair: Lists each attribute and its value as a compact key–value pair
    4. **Sentence**: Converts each attribute–value pair into a **natural language** sentence

Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Raw Clinical Metadata

Sentence
Sex is female, Age is 55, Anatomic Site General Challenge is anterior torso.

Proposed Framework

- **Multimodal indexing** and retrieval
  - **Build vector database** of image-metadata pairs to find semantically similar cases

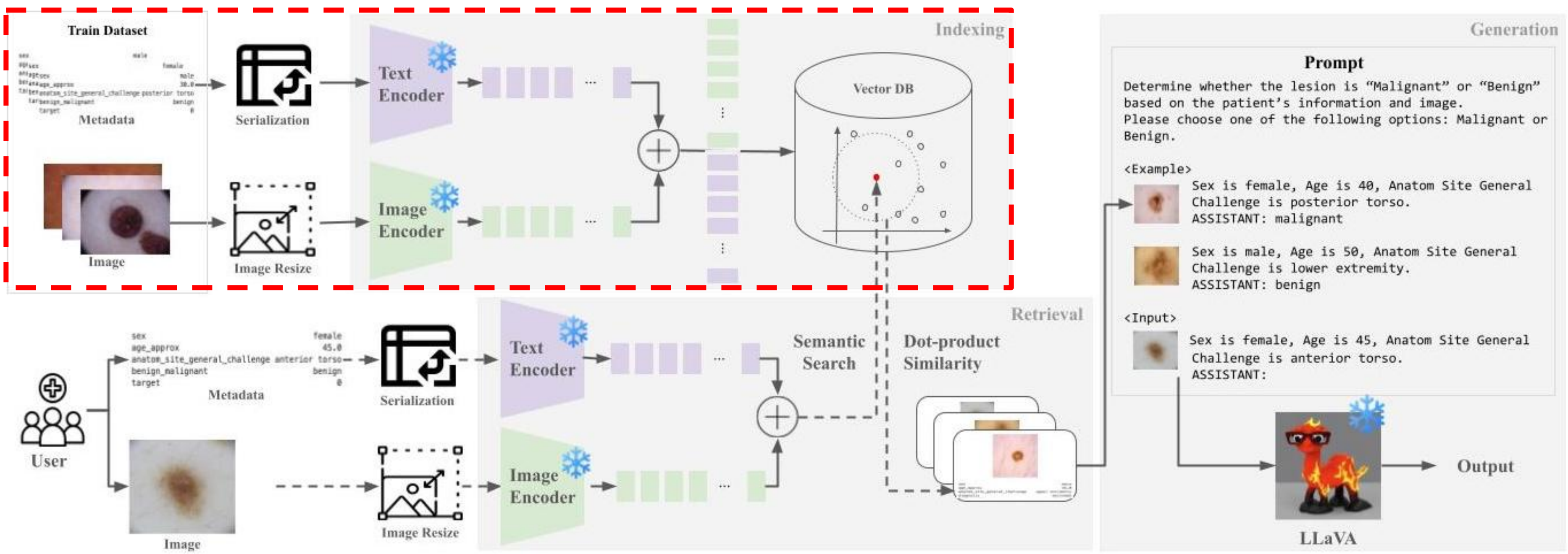
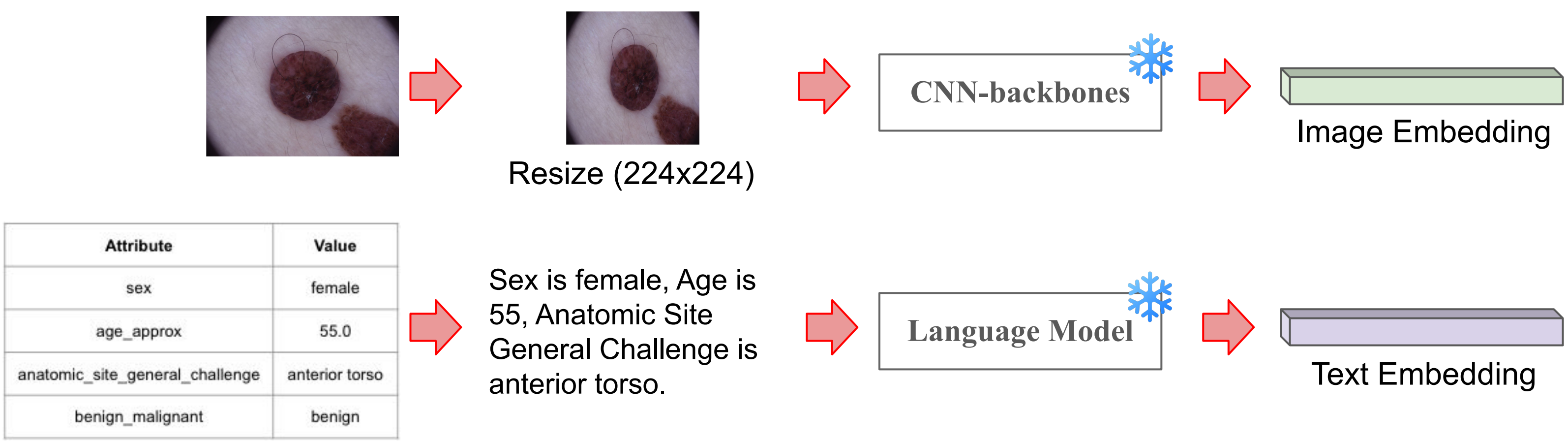


Figure 1. Proposed retrieval-augmented classification framework with sentence-based prompting.

Proposed Framework

- **Multimodal indexing** and retrieval
  - Each patient record is transformed into a **unified multimodal vector**:
    - **Image**: Resized to 224x224 and encoded using CNN backbones (ResNet [He et al., 2016], EfficientNet [Tan et al., 2021])
    - **Metadata**: Serialized into text and embedding using a pretrained language model (BERT [Devlin et al., 2019])

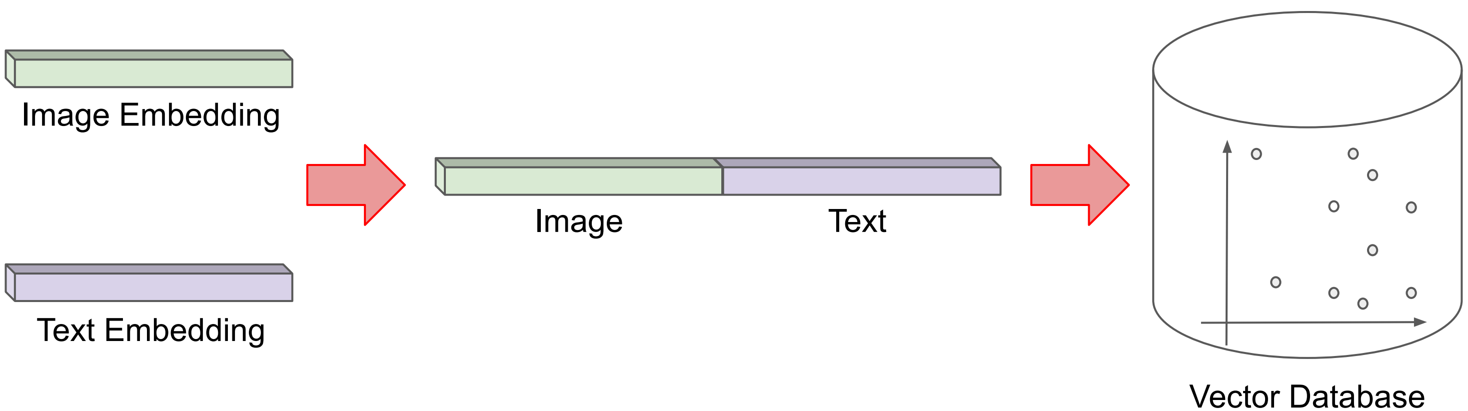


Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

He et al. (2016). Deep residual learning for image recognition.
Tan et al. (2021). Efficientnetv2: Smaller models and faster training.
Devlin et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Proposed Framework

- **Multimodal indexing** and retrieval
  - Each patient record is transformed into a **unified multimodal vector**
  - **Concatenated** vectors stored in **FAISS-based** database [Douze et al., 2024] for efficient similar nearest neighbor search



Douze et al. (2024). The faiss library.

Proposed Framework

- **Multimodal** indexing and **retrieval**
  - Build vector database of image-metadata pairs to **find** **semantically similar cases**

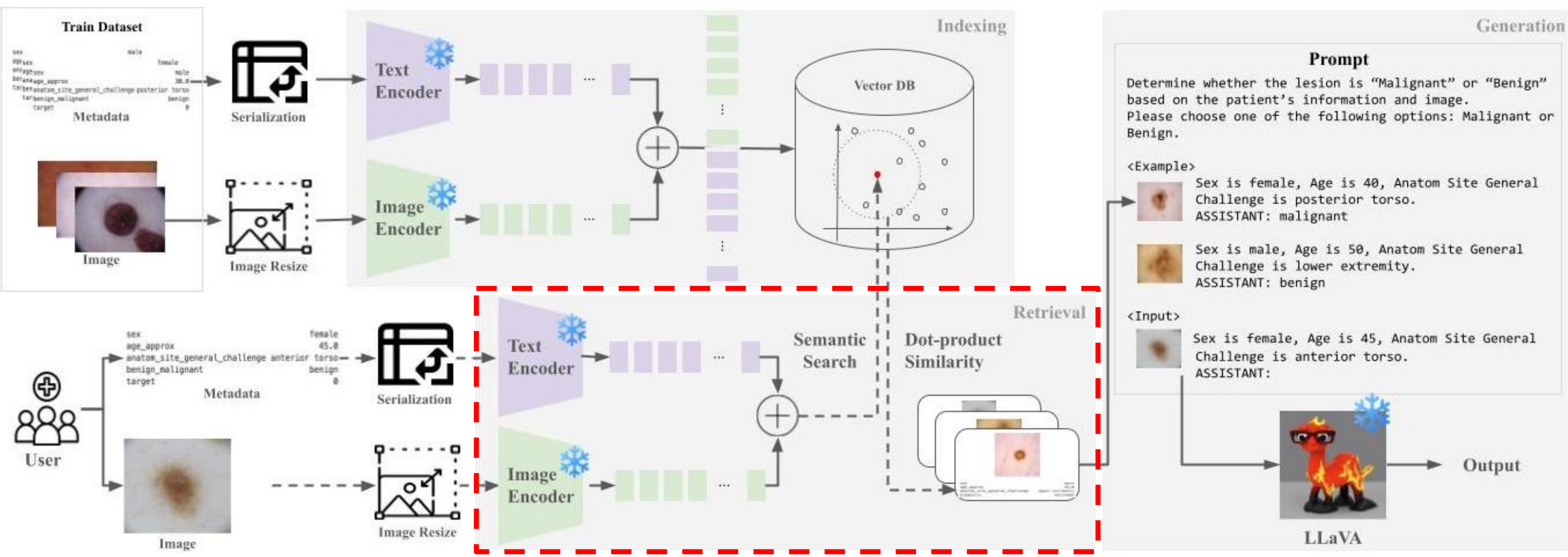
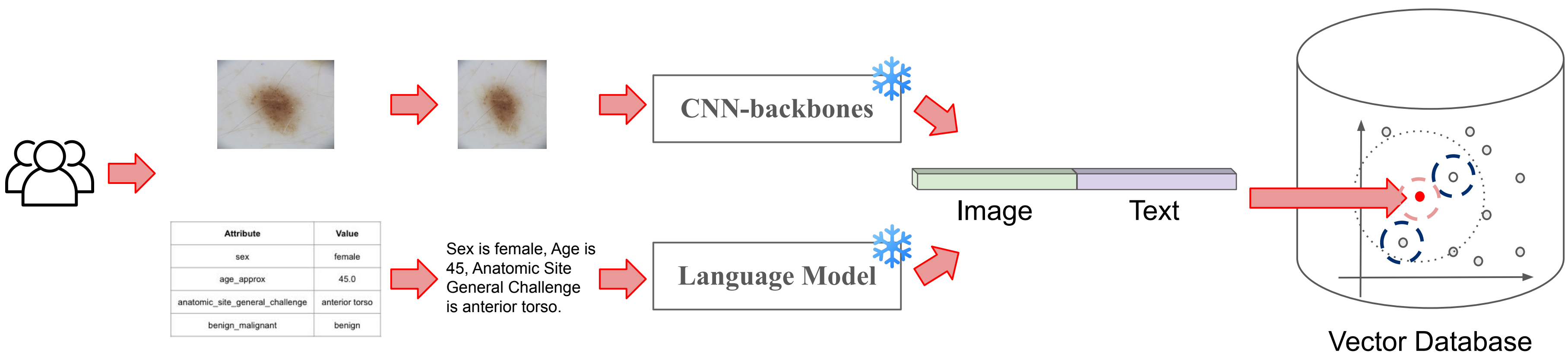


Figure 1. Proposed retrieval-augmented classification framework with sentence-based prompting.

Proposed Framework

- **Multimodal** indexing and **retrieval**
  - **Target query** (image & metadata) is encoded using the **same encoders**
  - Similarity between the query and stored vectors is computed using **dot-product**

➡ **Top-*K* (*K*-shot)** **most similar** patient cases retrieved as contextual examples



Experiment

- Experimental Setting
  - Dataset
    - **ISIC** (International Skin Imaging Collaboration) 2019 challenge dataset [1]
      - **Binary** classification task: Malignant vs. Benign
      - Data components: Dermoscopic images with corresponding patient metadata (age, sex, anatomical site)

Class	Train Size	Validation Size	Test Size	Total Size
Positive (Malignant)	3,137 (18.7%)	776 (18.5%)	1,695 (18.8%)	5,608 (18.7%)
Negative (Benign)	13,619 (81.3%)	3,414, (81.5%)	7,282 (81.2%)	24,315 (81.3%)

Table 1: Sample counts and class distribution (%) by split.

  - Evaluation Metrics
    - To evaluate the performance of melanoma classification,
      - **F1-score** as primary metric due to class imbalance and clinical importance

[1] Available at: https://challenge.isic-archive.com/data/#2019

Experiment

- Performance Comparison

	Image	Metadata	Model	Serialization	Accuracy	Balanced Accuracy	Precision	Sensitivity	F1	TN	TP	FN	FP
	✓	-	ResNet-50	-	0.6307	0.4920	0.1801	0.2690	0.2158	5206	456	1239	2076
	✓	-	ResNeXt-50	-	0.7380	0.5054	0.2022	0.1316	0.1594	6402	223	1472	880
Image-based	✓	-	EfficientNet-B0	-	0.6560	0.4913	0.1777	0.2265	0.1992	5505	384	1311	1777
	✓	-	EfficientNet-V2-S	-	0.6833	0.4959	0.1825	0.1947	0.1884	5804	330	1365	1478
	✓	-	EfficientNet-V2-M	-	0.6954	0.5061	0.1985	0.2018	0.2001	5901	342	1353	1381
	-	✓	Mistral 7B v1.0	html	0.5014	0.4885	0.1816	0.4678	0.2616	3708	793	902	3574
	-	✓	Mistral 7B v1.0	markdown	0.8034	0.5004	0.1983	0.0136	0.0254	7189	23	1672	93
	-	✓	Mistral 7B v1.0	attribute-value pair	0.7974	0.5143	0.3098	0.0596	0.1000	7057	101	1594	225
	-	✓	Mistral 7B v1.0	sentence	0.8129	0.5090	0.6364	0.0206	0.0400	7262	35	1660	20
	-	✓	Llama 3 8B Instruct	html	0.7906	0.4995	0.1843	0.0319	0.0543	7043	54	1641	239
Text-based	-	✓	Llama 3 8B Instruct	markdown	0.8104	0.5002	0.2308	0.0018	0.0035	7272	3	1692	10
	-	✓	Llama 3 8B Instruct	attribute-value pair	0.7986	0.4977	0.1491	0.0142	0.0259	7145	24	1671	137
	-	✓	Llama 3 8B Instruct	sentence	0.7916	0.4988	0.1765	0.0283	0.0488	7058	48	1647	224
	-	✓	Vicuna 7B v1.5	html	0.6547	0.4873	0.1725	0.2183	0.1927	5507	370	1325	1775
	-	✓	Vicuna 7B v1.5	markdown	0.6930	0.5023	0.1925	0.1959	0.1942	5889	332	1363	1393
	-	✓	Vicuna 7B v1.5	attribute-value pair	0.7037	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
	✓	✓	BERT + ResNet-50	html	0.6519	0.5000	0.1889	0.2560	0.2174	5418	434	1261	1864
	✓	✓	BERT + ResNet-50	markdown	0.6474	0.4971	0.1854	0.2555	0.2148	5379	433	1262	1903
	✓	✓	BERT + ResNet-50	attribute-value pair	0.7030	0.5017	0.1917	0.1782	0.1847	6009	302	1393	1273
	✓	✓	BERT + ResNet-50	sentence	0.7016	0.5042	0.1959	0.1870	0.1914	5981	317	1378	1301
	✓	✓	BERT + ResNeXt-50	html	0.6819	0.5079	0.2000	0.2283	0.2132	5734	387	1308	1548
	✓	✓	BERT + ResNeXt-50	markdown	0.7152	0.5105	0.2084	0.1817	0.1941	6112	308	1387	1170
	✓	✓	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-B0	html	0.6858	0.4969	0.1841	0.1935	0.1887	5828	328	1367	1454
	✓	✓	BERT + EfficientNet-B0	markdown	0.7020	0.5031	0.1941	0.1835	0.1887	5991	311	1384	1291
Early-Fusion	✓	✓	BERT + EfficientNet-B0	attribute-value pair	0.7083	0.5108	0.2076	0.1935	0.2003	6030	328	1367	1252
	✓	✓	BERT + EfficientNet-B0	sentence	0.6891	0.5012	0.1907	0.1994	0.1950	5848	338	1357	1434
	✓	✓	BERT + EfficientNet-V2-S	html	0.7062	0.5030	0.1942	0.1764	0.1849	6041	299	1396	1241
	✓	✓	BERT + EfficientNet-V2-S	markdown	0.7027	0.5103	0.2059	0.2012	0.2035	5967	341	1354	1315
	✓	✓	BERT + EfficientNet-V2-S	attribute-value pair	0.6891	0.5024	0.1925	0.2024	0.1973	5843	343	1352	1439
	✓	✓	BERT + EfficientNet-V2-S	sentence	0.6987	0.5022	0.1924	0.1864	0.1894	5956	316	1379	1326
	✓	✓	BERT + EfficientNet-V2-M	html	0.7024	0.4967	0.1830	0.1664	0.1743	6023	282	1413	1259
	✓	✓	BERT + EfficientNet-V2-M	markdown	0.7105	0.5011	0.1908	0.1646	0.1768	6099	279	1416	1183
	✓	✓	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
	✓	✓	LLaVa 7B v1.5 hf	html	0.5845	0.6113	0.2608	0.6543	0.3729	4138	1109	586	3144
VLM with	✓	✓	LLaVa 7B v1.5 hf	markdwon	0.6915	0.6003	0.2944	0.4537	0.3571	5439	769	926	1843
Zero-shot	✓	✓	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
	✓	✓	BERT + ResNet-50	html	0.7400	0.7223	0.3932	0.6938	0.5019	5467	1176	519	1815
	✓	✓	BERT + ResNet-50	markdown	0.8168	0.7619	0.5112	0.6737	0.5813	6190	1142	553	1092
	✓	✓	BERT + ResNet-50	attribute-value pair	0.8787	0.7858	0.6952	0.6366	0.6646	6809	1079	616	473
	✓	✓	BERT + ResNet-50	sentence	0.8722	0.7775	0.6743	0.6254	0.6489	6770	1060	635	512
	✓	✓	BERT + ResNeXt-50	html	0.7396	0.7202	0.3921	0.6891	0.4998	5471	1168	527	1811
	✓	✓	BERT + ResNeXt-50	markdown	0.8268	0.7774	0.5314	0.6979	0.6034	6239	1183	512	1043
	✓	✓	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-B0	html	0.7459	0.7300	0.4015	0.7044	0.5115	5502	1194	501	1780
Ours	✓	✓	BERT + EfficientNet-B0	markdown	0.8166	0.7654	0.5108	0.6832	0.5846	6173	1158	537	1109
($k = 2$)	✓	✓	BERT + EfficientNet-B0	attribute-value pair	0.8847	0.8004	0.7070	0.6649	0.6853	6815	1127	568	467
	✓	✓	BERT + EfficientNet-B0	sentence	0.8791	0.7911	0.6916	0.6496	0.6699	6791	1101	594	491
	✓	✓	BERT + EfficientNet-V2-S	html	0.7195	0.6897	0.3628	0.6419	0.4636	5371	1088	607	1911
	✓	✓	BERT + EfficientNet-V2-S	markdown	0.7959	0.7190	0.4682	0.5953	0.5242	6136	1009	686	1146
	✓	✓	BERT + EfficientNet-V2-S	attribute-value pair	0.8605	0.7553	0.6426	0.5863	0.6132	1452	214	151	119
	✓	✓	BERT + EfficientNet-V2-S	sentence	0.8559	0.7464	0.6308	0.5705	0.5991	6716	967	728	566
	✓	✓	BERT + EfficientNet-V2-M	html	0.7123	0.6746	0.3505	0.6142	0.4463	5353	1041	654	1929
	✓	✓	BERT + EfficientNet-V2-M	markdown	0.7881	0.7080	0.4523	0.5794	0.5080	6093	982	713	1189
	✓	✓	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	0.5422	0.5706	6675	933	762	593
	✓	✓	BERT + ResNet-50	html	0.8066	0.7821	0.4920	0.7428	0.5919	5982	1259	436	1300
	✓	✓	BERT + ResNet-50	markdown	0.7838	0.7796	0.4571	0.7729	0.5744	5726	1310	385	1556
	✓	✓	BERT + ResNet-50	attribute-value pair	0.8456	0.7745	0.5801	0.6602	0.6175	6472	1119	576	810
	✓	✓	BERT + ResNet-50	sentence	0.8574	0.7824	0.6134	0.6619	0.6368	6575	1122	573	707
	✓	✓	BERT + ResNeXt-50	html	0.8200	0.7974	0.5158	0.7611	0.6149	6071	1290	405	1211
	✓	✓	BERT + ResNeXt-50	markdown	0.7941	0.7971	0.4734	0.8018	0.5953	5770	1359	336	1512
	✓	✓	BERT + ResNeXt-50	attribute-value pair	0.8479	0.7841	0.5833	0.6814	0.6286	6457	1155	540	825
	✓	✓	BERT + ResNeXt-50	sentence	0.8703	0.8006	0.6473	0.6885	0.6672	6646	1167	528	636
	✓	✓	BERT + EfficientNet-B0	html	0.8135	0.7911	0.5041	0.7552	0.6046	6023	1280	415	1259
Ours	✓	✓	BERT + EfficientNet-B0	markdown	0.7907	0.7915	0.4680	0.7929	0.5886	5754	1344	351	1528
($k = 4$)	✓	✓	BERT + EfficientNet-B0	attribute-value pair	0.8448	0.7821	0.5752	0.6814	0.6238	6429	1155	540	853
	✓	✓	BERT + EfficientNet-B0	sentence	0.8640	0.7978	0.6267	0.6914	0.6575	6584	1172	523	698
	✓	✓	BERT + EfficientNet-V2-S	html	0.7833	0.7415	0.4507	0.6743	0.5403	5889	1143	552	1393
	✓	✓	BERT + EfficientNet-V2-S	markdown	0.7609	0.7492	0.4230	0.7304	0.5357	5593	1238	457	1689
	✓	✓	BERT + EfficientNet-V2-S	attribute-value pair	0.8220	0.7371	0.5250	0.6006	0.5603	6361	1018	677	921
	✓	✓	BERT + EfficientNet-V2-S	sentence	0.8397	0.7552	0.5694	0.6195	0.5934	6488	1050	645	794
	✓	✓	BERT + EfficientNet-V2-M	html	0.7734	0.7252	0.4331	0.6478	0.5191	5845	1098	597	1437
	✓	✓	BERT + EfficientNet-V2-M	markdown	0.7425	0.7272	0.3971	0.7027	0.5075	5474	1191	504	1808
	✓	✓	BERT + EfficientNet-V2-M	attribute-value pair	0.8047	0.7092	0.4851	0.5558	0.5180	6282	942	753	1000
	✓	✓	BERT + EfficientNet-V2-M	sentence	0.8298	0.7353	0.5461	0.5835	0.5642	6460	989	706	822

Experiment

- Can **clinical metadata** enhance image-based melanoma diagnosis?
  - We Found
    - Clinical **metadata** provides **powerful diagnostic** cues beyond what images reveal
    - Relying solely on images **overlooks crucial clinical indicators**
    - **Integrating clinical context** is essential for reliable melanoma classification

Fine-Tuned	Modality: Image	Modality: Metadata	Model	Serialization	Accuracy	F1 Score
✓	✓	-	ResNet 50	-	0.6307	0.2158
✓	✓	-	ResNeXt 50	-	0.7380	0.1594
✓	✓	-	EfficientNet B0	-	0.6560	0.1992
✓	✓	-	EfficientNet V2 S	-	0.6833	0.1884
✓	✓	-	EfficientNet V2 M	-	0.6954	0.2001
-	-	✓	Mistral 7B v1.0	HTML	0.5014	0.2616 +0.0458
-	-	✓	Vicuna 7B v1.5	Markdown	0.6930	0.1942
-	-	✓	Vicuna 7B v1.5	Attribute-Value pair	0.7032	0.2352
-	-	✓	Vicuna 7B v1.5	Sentence	0.6063	0.2613

Experiment

- Can a **VLM effectively process** dermoscopic images for diagnostic classification?
  - We Found
    - **Zero-shot VLMs outperform** multimodal embedding-level (early fusion) methods
    - Pretrained models achieve **~71.5% F1 improvement** without requiring additional fine-tuning
    - **Effective joint processing** of dermoscopic images and clinical metadata

	Modality						
	Image	Metadata	Model	Serialization	Accuracy	F1 Score	
Early Fusion	✓	✓	BERT + ResNet-50	HTML	0.6519	0.2174	
	✓	✓	BERT + ResNet-50	Markdown	0.6474	0.2148	
	✓	✓	BERT + EfficientNet-B0	Attribute-Value pair	0.7083	0.2003	
	✓	✓	BERT + EfficientNet-B0	Sentence	0.6891	0.1950	
Zero-Shot VLM	✓	✓	LLaVA 7B v1.5 hf	HTML	0.5845	0.3729	
	✓	✓		Markdown	0.6915	0.3581	+ 71.5%
	✓	✓		Attribute-Value pair	0.7126	0.3729	
	✓	✓		Sentence	0.5610	0.3303	

Experiment

- Can a VLM achieve **clinically acceptable** diagnostic accuracy?
  - We Found
    - Zero-shot VLMs **outperform** baseline methods
      - Confirming the **efficacy of joint processing** of dermoscopic images and clinical metadata
    - Performance improved even without fine-tuning, showing **potential for generalization**
    - **F1 score remains below** the threshold for reliable clinical application
      - The need for further refinement

Modality					
Image	**Metadata**	**Model**	**Serialization**	**Accuracy**	**F1 Score**
✓	-	ResNet 50	-	0.6307	0.2158
-	✓	Mistral 7B v1.0	HTML	0.5014	0.2616
✓	✓	BERT + ResNet-50	HTML	0.6519	0.2174
✓	✓	LLaVA 7B v1.5 hf (0-shot)	Attribute-Value pair	0.7126	0.3729

Experiment

- Does **RAG** enhance performance by refining clinical cases **without fine tuning**?
  - We Found
    - RAG substantially **improves** F1 score without fine-tuning
    - Best performance at **2-shot** (Top-2) retrieval
    - Providing relevant clinical cases **strengthens diagnostic reasoning capabilities**

Top-*K* (*K*-Shot)	Accuracy	F1 Score
0-Shot	0.7126	0.3729
1-Shot	0.8833	0.6381
2-Shot	0.8876	0.6864 + 84.0%
3-Shot	0.8694	0.6648
4-Shot	0.8479	0.6286

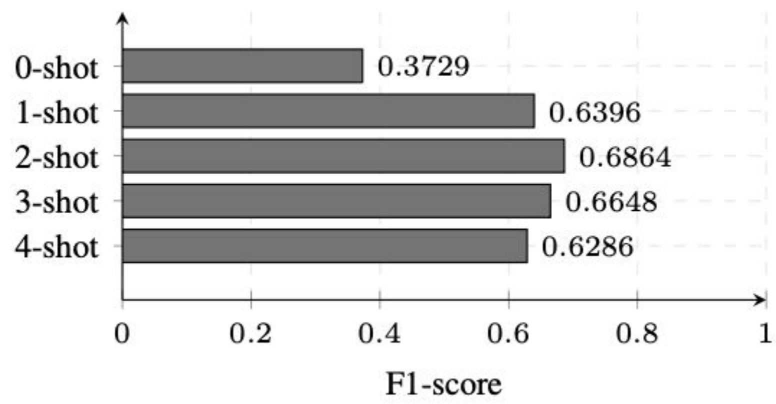


Figure 2: Effect of retrieval count *K*-Shot (Top-*K*) on performance using **BERT + ResNeXT-50** and **attribute-value pair** format.

Conclusion

- Proposed a **retrieval-augmented VLM framework** for melanoma classification
  - Incorporates **semantically similar cases** to enhance diagnostic context
- **Outperformed** all baselines, especially under zero-shot constraints
  - Without fine-tuning — making it practical for real-world clinical workflows
- Shows potential for broader use in multimodal medical AI applications
- Limitations
  - **Dependence** on curated training data and need to improve **retrieval speed** for real-time use

References

- S. N. Markovic, L. A. Erickson, R. D. Rao, R. R. McWilliams, L. A. Kottschade, E. T. Creagan, R. H. Weenig, J. L. Hand, M. R. Pittelkow, B. A. Pockaj, et al. Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis. In Mayo clinic proceedings, volume 82, pages 364–380. Elsevier, 2007.
- A. F. Duarte, B. Sousa-Pinto, L. F. Azevedo, A. M. Barros, S. Puig, J. Malvehy, E. Haneke, and O. Correia. Clinical abcde rule for early melanoma detection. European Journal of Dermatology, 31(6):771–778, 2021.
- T. Alafghani. A CMOS 10-bit SAR ADC, with on-chip offset cancellation, for near-field, mm-wave imaging technique, applied to skin cancer detection. In Doctoral dissertation, Masdar Institute of Science and Technology, 2018.
- A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639):115–118, 2017.
- S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. 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.
- T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. Von Kalle. Skin cancer classification using convolutional neural networks: systematic review. Journal of medical Internet research, 20(10):e11936, 2018.
- Y. Wang, Y. Feng, L. Zhang, J. T. Zhou, Y. Liu, R. S. M. Goh, and L. Zhen. Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images. Medical Image Analysis, 81, 2022.
- M. Akrout, K. D. Cirone, and R. Vender. 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.
- B. Zheng, B. Gou, J. Kil, H. Sun, and Y. Su. Gpt-4v(ision) is a generalist web agent, if grounded. arXiv preprint arXiv:2401.01614, 2024.
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.
- N. Shifai, R. van Doorn, J. Malvehy, and T. E. Sangers. Can chatgpt vision diagnose melanoma? an exploratory diagnostic accuracy study. Journal of the American Academy of Dermatology, 90(5):1057–1059, 2024.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
- M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In International conference on machine learning, pages 10096–10106. PMLR, 2021.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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), pages 4171–4186, 2019.
- M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.E. Mazar´ e, M. Lomeli, L. Hosseini, and H. J´ egou. The faiss library. arXiv preprint arXiv:2401.08281, 2024.
- 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, 5(1):1–9, 2018.
- N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, 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. IEEE, 2018.
- M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, et al. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288, 2019.
- S. Xie, R. Girshick, P. Doll´ ar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.
- A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623, 2023.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763, 2021. PMLR.

HAIL

HANDONG ARTIFICIAL INTELLIGENCE LAB

Thank you for your attention

Title	**Multimodal Clinical Decision Support for Melanoma Diagnosis Using Retrieval-Augmented Generation and Vision-Language Models**

Presenter	Jihyun Moon (jhmoon@handong.ac.kr)
Advisor	Charmgil Hong (charmgil@handong.ac.kr)