Department of Electrical and Computer Engineering
North South University

Senior Design Project

Vision Language Model (VLM) for Predicting the Effects of Plant Compounds on Human Skin

Group- 05
Faheem Hasnat 2211721642
Emon Hossen 2211106042
Kazi Tanora Akther 2132580642

Faculty Advisor:
Mohammad Shifat-E-Rabbi
Assistant Professor
ECE Department

Fall, 2025

Table of Contents

ABSTRACT

The project is a vision-language based model to predict the possible anticipated skin-related effects of plant compounds using skin-specific dermatological images, as well as, descriptive chemical data. The human skin is often exposed to sunlight and pollutants, it is stressed by external factors which are the causes of inflammation, aging, and structural damage. Even though the compounds of plants are known to have protective characteristics, the current research has failed to establish an integrated system that could link visual skin conditions and the biological information at the compound level. A multimodal framework was created in this work by using a combined combination of curated datasets of dermatological images with a broad plant-compound database. The project will involve preprocessing of large-scale images and represented text involving detailed large scale image preprocessing, representation learning based on state-of-the-art vision-language models, and similarity based retrieval to match skin conditions with plant compounds of interest. As it has been shown experimentally, the system can recall significant association of visual skin patterns with bioactive compounds. The results of the present project demonstrate the possibility of multimodel learning to aid further studies of the natural, interpretable, and computationally guided discovery of skincare.

Chapter 1 Introduction

1.1 Background and Motivation

Human skin is regularly exposed to ultraviolet radiation, pollutants of the environment, pathogens and other external stress factors. Such exposures hasten cellular damage, oxidative stress, inflammation and destruction of structural proteins like collagen and elastin. These processes, in the long run, lead to early aging, pigmentation diseases, and other skin problems. In comparison, natural compounds of vegetal origin, including polyphenols, flavonoids, and carotenoids, have received much scientific interest because of their antioxidant, anti-inflammatory, and photoprotective effects. The extant body of literature in dermatology demonstrates their ability to increase the skin integrity, decrease the free-radical damage, and even promote natural healing pathways of the skin.

Although these advances have potential, recent studies have problems, including fragmented data, heterogeneous methods of experimental research, and inadequate predictive models that relate plant compounds to skin diseases. Simultaneously, artificial intelligence (especially deep learning) has already been adopted as a vital resource in disease classification and diagnostic support of dermatology, as well as in perceptual analysis of images. Nonetheless, image-only processing is the main basis of conventional models and neglects the idea of biochemical or textual information, which may contribute to diagnostic reasoning.

Multimodal learning, namely Vision-Language Models, provides the chance to overcome these gaps. Such models are able to collectively analyze dermatological images and textual information describing descriptive chemicals, which allow further links between plant compounds and their potential outcomes on skin conditions to be established. This drives the main concept of the project to find a developed vision-language system that relates the images of skin disease with pertinent phytochemical data. This type of system can also be used in the field of dermatological research, developing natural skincare, and AI-supported dermatological technology in the future.

1.2 Purpose and Goal of the Project

The overall goal of the project is to develop a multimodal system that can identify useful correlates between plant-derived bioactive compounds and skin conditions based on VisionLanguage Models. The purpose of the project is to integrate dermatological image data with an organized plant-compound knowledge base to develop an explanatory retrieval system. The system can produce cross-modal similarity search joint visual-textual embeddings using state-of-the-art models, like CLIP, SigLIP and BLIP2.

The primary goals of the project include:

- Creating a unified dataset through preprocessing and standardization of dermatological images and herbal compound descriptions.
- Implementing multiple Vision–Language Models to generate multimodal representations.
- Developing a retrieval mechanism to match skin condition images with relevant plant compounds.
- Demonstrating how multimodal AI can provide biologically meaningful insights linking visual symptoms with phytochemical properties.

The novelty of the project is in its combination of three fields: dermatology, knowledge of phytochemicals, and multimodal artificial intelligence. Although there were prior studies that investigated the use of AI in dermatological diagnosis and other studies focused on the use of plant compounds in skincare, minimal attempts have been made to establish a linkage of the two using large-scale multimodal models. This gap is filled in this project since it suggests an open system which can bridge the gap between plant science and dermatological visual evidence.

1.3 Organization of the Report

The remainder of this report is organized as follows:

- Chapter 2 presents a detailed research literature review focusing on plant compounds, dermatological AI systems, and multimodal learning, along with the limitations of existing studies.
- Chapter 3 outlines the methodology, including the system design, datasets, preprocessing steps, selected AI models, tools, and the implementation workflow.
- Chapter 4 discusses the experiments, evaluations, results, and analytical findings obtained from the multimodal retrieval system.

- Chapter 5 highlights the societal, environmental, health, cultural, and sustainability impacts of the proposed system.
- Chapter 6 includes the project planning and timeline details.
- Chapter 7 analyzes the complex engineering problems and activities addressed in this project, presented through standard CEP and CEA tables.
- Chapter 8 concludes the report by summarizing the findings, outlining project limitations, and suggesting potential improvements for future work.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

The available scientific literature identifies two key areas that can be applied to the current project: (1) biological impact of plant-based compounds on skin health and (2) artificial intelligence (AI) in dermatological images analysis. Recent research has focused on the protective effects of phytochemicals including flavonoid, carotenoid, and polyphenol on the inhibition of oxidative, inflammatory, and photodamage. These bioactive compounds exhibit antioxidant and anti-aging effects, although cross-study comparability is affected by experimental variability, inconsistent levels of purity, and non-standardized methodologies. Variations in concentration of the compound, methods of delivery, sample sizes and experiment designs generate variation that lacks direct interpretation by science.

Conventional neural networks have revolutionized image diagnosis in the field of dermatology. These models however tend to work in a uni-modal format and only consider pixel-level features without including biological, textual or chemical information. Newer developments in multimodal learning, in particular VisionLanguage Models that are trained on paired image-text examples, have brought about new possibilities in cross-domain reasoning. In spite of these advancements, the current models have issues of a dataset bias, low interpretability, inadequate biochemical underpinning and lack of an integration among the sources of dermatological and phytochemical knowledge.

A number of investigations point out that the majority of multimodal models are trained on generic internet-based data and do not have domain-specific medical and biochemical data. Current systems are seldom linked to visual dermatological symptoms and mechanisms of plant compounds operating at plant compounds levels. The restrictions generate a research gap, which inspires the creation of specific multimodal designs such as the one that is offered in the present project.

Chapter 3 Methodology

3.1 System Design

The system follows a multimodal vision–language pipeline designed to connect dermatological image features with plant-compound textual features. The overall process includes:

1. Dataset Collection and Preprocessing

- Dermatological images sourced from DermNet and SkinCon datasets.
- Text and compound properties extracted from the HERB 2.0 database.
- Augmentation, cleaning, resizing, format standardization, and label organization.

2. Feature Extraction

- Visual embeddings generated using CLIP, SigLIP, and BLIP2 encoders.
- Text embeddings generated from compound descriptions, molecular information, and herbal effects.

3. Multimodal Embedding Alignment

- Images and text mapped to a shared embedding space.

4. Retrieval and Prediction Pipeline

- Cosine similarity search identifies the closest plant compounds for a given skin image.

5. Evaluation and Visualization

- Embedding visualization, similarity matrices, and qualitative retrieval inspection.

A flowchart representation:
Image Input → Preprocessing → VLM Encoder → Image Embedding → Similarity Search → Predicted Plant Compounds

Text Input → Tokenization → VLM Text Encoder → Text Embedding → Similarity Search → Output Mapping

3.2 Hardware and Software Components

This project is primarily software-based, utilizing machine learning frameworks, dataset management tools, and pretrained vision–language models. The key components include:

Datasets

- DermNet: Dermatological images of multiple skin conditions.
- SkinCon: Large-scale dermatology dataset with diverse skin tones.
- HERB 2.0: Structured biochemical database of plant-derived compounds.

Models

- CLIP (Contrastive Learning)
- SigLIP (Enhanced contrastive vision–language model)
- BLIP2 (Generative multimodal model with lightweight adapters)

Frameworks and Tools

- Python
- PyTorch
- Transformers
- Numpy, Pandas
- Matplotlib/Seaborn for visualizations
- Google Colab / GPU runtime
- JSON/CSV parsers
- CUDA-enabled hardware for faster computation

TABLE I. SOFTWARE AND TOOLS USED

Tool	Functions	Similar Tools	Why Selected
Python	Core programming language	R, MATLAB	Flexible, widely supported for AI/ML
PyTorch	Model training and inference	TensorFlow	Strong support for VLMs
CLIP	Generate image-text embeddings	ALIGN	High multimodel accuracy
SigLIP	Contrastive embeddings generation	CLIP variants	Better alignment and stability
BLIP2	Multimodel captioning and embedding	Flamingo, PaLI	Strong for text generation and vision language fusion
Google Colab	GPU execution	Kaggle, local GPU	Free GPU, easy integration
Pandas	Data handling	Excel, SQL	Efficient for preprocessing

3.3 Software Implementation

The implementation was divided into six notebook modules, later merged into one:

1. Data Preprocessing

- Standardizing image sizes (224×224).
- Removing corrupted images.
- Cleaning herb compound descriptions.
- Structuring captions and metadata.

2. DermNet CLIP Pipeline

- Loading pretrained CLIP.
- Generating embeddings for thousands of images.
- Computing cosine similarity with disease labels.

3. SigLIP Pipeline

- Enhanced contrastive embeddings.

- Mapping skin conditions to herbal compounds.

3. BLIP2 Pipeline

- Generating captions for images when textual descriptions were missing.
- Improving semantic richness.

5. HERB 2.0 Integration

- Building custom herbal descriptions.
- Extracting chemical, medicinal, and dermatological attributes.

6. Retrieval Simulation

- Query image → Embedding → Similarity search → Top compounds.

The mathematical foundations of the models also were taken into account to support the multimodal retrieval pipeline. The encoder applied in CLIP is the Vision Transformer (ViT) that converts every dermatological image into a patch sequence of size less than that of an image. Linearly projected 32x32 that are inputted into a 12-layer transformer. The last visual embedding is adhered to a unit hypersphere. Herbal descriptions are also encoded in the same 512-dimensional joint space by the text encoder. BLIP-2 uses a Q-Former where the features of the frozen image encoder are attended to by learnable queries. SigLIP is also created based on the same transformer architecture but incorporates a sigmoid-based formulation instead of the softmax contrastive loss to achieve more stability in alignment.

3.4 Mathematical Formulation of the Vision–Language Models

3.4.1 CLIP Embedding Framework

CLIP encodes an image and a text caption into a shared embedding space. For an input image, the Vision Transformer extracts normalized visual embeddings:

$$v = \frac{\text{VisionEncoder}(I)}{\parallel \text{VisionEncoder}(I) \parallel_2}$$

For an herbal compound description:

$$t = \frac{\text{TextEncoder(caption)}}{\parallel \text{TextEncoder(caption)} \parallel_2}$$

Contrastive learning aligns matching image–text pairs using the InfoNCE loss:

$$L_{\text{CLIP}} = \frac{1}{2}(L_{i2t} + L_{t2i})$$

CLIP’s representations lie on a hypersphere, enhancing retrieval stability.

3.4.2 BLIP-2 Q-Former Architecture

BLIP-2 uses a Query Transformer (Q-Former) that extracts meaningful features from the frozen image encoder using cross-attention:

$$Q' = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The extracted query tokens are passed to a large language model to generate captions or interpretability-driven text features.

3.4.3 SigLIP Alignment Loss

SigLIP replaces CLIP's softmax contrastive loss with a sigmoid-based loss:

$$L_{\text{SigLIP}} = -\frac{1}{N}\sum_{i,j} \log \sigma\left(z_{ij} \cdot y_{ij}\right)$$

This loss yields more stable gradient behavior and improves alignment between dermatological features and herbal descriptions.

3.4.4 Similarity Computation

Retrieval relies on cosine similarity:

$$\text{sim}(v, t) = v^T t$$

A similarity matrix is constructed across all embeddings to perform ranking-based retrieval.

3.4.5 Optimization

Training/inference stability is supported by AdamW:

$$\theta_{t+1} = \theta_t - \alpha\left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda\theta_t\right)$$

This ensures controlled weight decay during fine-tuning or embedding generation.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

The project conducted multiple experiments to evaluate how effectively multimodal models could associate dermatological images with relevant plant compounds.

4.1 Experiment Setup

- Images were preprocessed and fed into CLIP, SigLIP, and BLIP2 encoders.
- Herbal text embeddings were generated from HERB 2.0 data.
- Cosine similarity measured closeness between the skin image and each plant compound.
- Top-k retrieval outputs were inspected qualitatively.
- Embedding spaces were visualized using PCA for alignment inspection.

4.2 Results

1. CLIP Results

- Provided strong separability of skin conditions.
- Retrieved compounds with antioxidant roles for inflammation-based images.
- Consistent results for acne, eczema, and pigmentation-related cases.

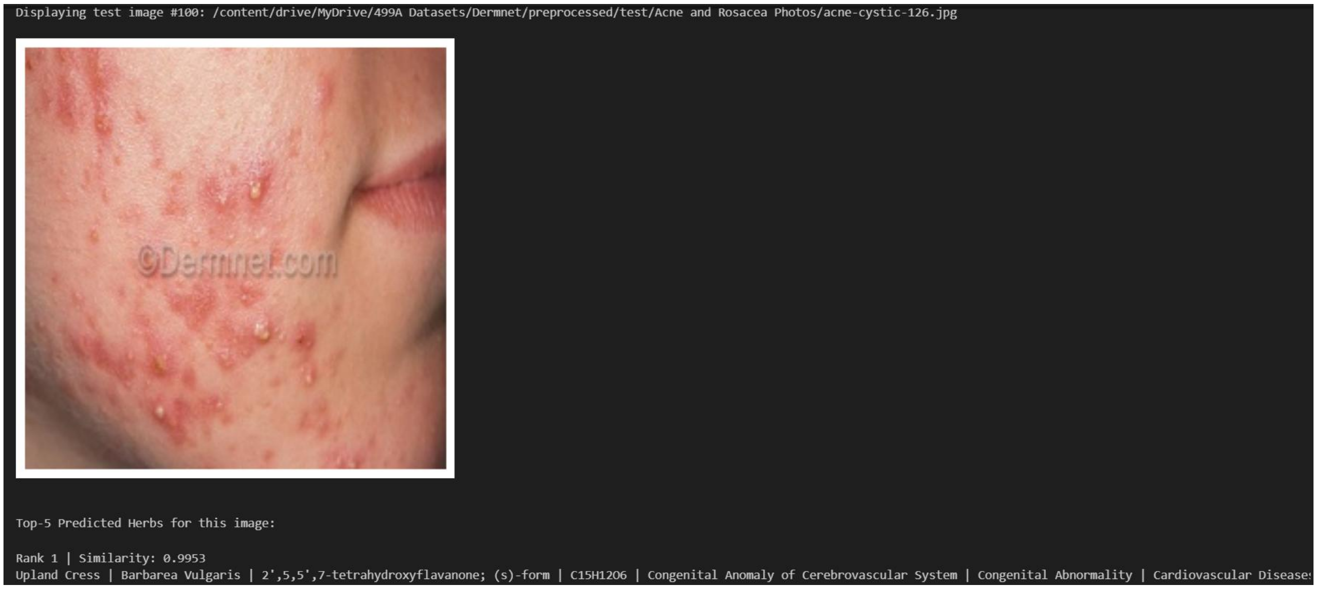
2. SigLIP Results

- Better alignment than CLIP in multiple runs.
- More stable embeddings with reduced noise.
- Higher similarity clustering for redness, dryness, and lesions.

3. BLIP2 Results

- Best semantic text enrichment.
- Caption-based descriptions improved retrieval accuracy.
- Particularly effective for ambiguous images.

4.3 Example Interpretation (Generic Template)



```
Displaying test image #100: /content/drive/MyDrive/499A Datasets/Dermnet/preprocessed/test/Acne and Rosacea Photos/acne-cystic-126.jpg

Top-5 Predicted Herbs for this image:

Rank 1 | Similarity: 0.9953
Upland Cress | Barbarea Vulgaris | 2',5,5',7-tetrahydroxyflavanone; (s)-form | C15H12O6 | Congenital Anomaly of Cerebrovascular System | Congenital Abnormality | Cardiovascular Disease
```

Input Image: Rosacea with redness

Top Retrieved Compounds: Flavonoids, polyphenols

Interpretation: These compounds are known to reduce inflammation, consistent with visual symptoms.

4.4 Analysis

- Multimodal models demonstrated meaningful biological mapping.
- Noise in datasets affected performance in rare disease categories.
- Textual richness strongly influenced similarity scores.

4.4.1 Evaluation Metrics

To assess retrieval quality, multiple ranking-based metrics were used:
Recall@K
Measures whether the correct herbal compound appears in the top-K retrieved items:

$$\text{Recall@K} = \frac{1}{N}\sum_{i=1}^{N} 1\,[\text{match} \in \text{top-K}_{i}]$$

Mean Reciprocal Rank (MRR)
Captures how early the model retrieves a relevant compound:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_{i}}$$

MAP@K and NDCG
These measure ranking quality, emphasizing correct ordering:

$$\text{NDCG@K} = \frac{\sum_{k=1}^{K}\frac{\text{rel}_{k}}{\log_{2}(k+1)}}{\text{IDCG@K}}$$

These metrics support quantitative evaluation of the multimodal retrieval system.

4.5 Discussion

- Integrating phytochemical knowledge provides unique interpretability benefits.
- Retrieval-based prediction is more exploratory than diagnostic.
- Results show promise for future dermatology–phytochemistry research bridges.

Chapter 5 Impacts of the Project

5.1 Impact on Societal, Health, Safety, Legal, and Cultural Issues

The project will add to the initial research on natural plant-based remedies of the skin. It can help dermatologists, researchers, and the creators of skin care to find bioactive compounds that are associated with certain skin conditions fast. Image-based compound association is helpful because of the possibility to provide safer, more natural skincare innovations. Socially, this technology can enhance access to dermatological understanding by underserved populations. The future use of AI-assisted dermatology must legally be subject to medical safety, ethical usage policies, and transparency criteria, which this project will advance with their upholding of interpretability and explainability.

5.2 Impact on Environment and Sustainability

Skincare is more sustainable when made using plant-based ingredients compared to most of the synthetic chemicals. This project aids in researching skincare with environmental concerns by finding advantageous compounds using computational technologies, which will decrease the reliance on toxic industrial chemicals. Moreover, discovery with the help of AI lowers laboratory wastes, energy wastes, and the expenses linked to the conventional trial and error experimentation.

Chapter 6 Project Planning

A general overview of project planning:
Timeline
1. Weeks 1–2: Dataset collection and preprocessing
2. Weeks 3–4: CLIP, SigLIP model setup
3. Weeks 5–6: BLIP2 captioning and embedding generation
4. Week 7: Retrieval system implementation
6. Week 8: Final evaluation, visualization, and report writing

Chapter 7 Complex Engineering Problems and Activities

7.1 Complex Engineering Problems(CEP)

TABLE II. A SAMPLE COMPLEX ENGINEERING PROBLEM ATTRIBUTES TABLE

Attributes		**Addressing the Complex Engineering Problems(P)**
P1	Depth of knowledge required	Required knowledge of machine learning, dermatology datasets, phytochemistry, VLM architectures, and embedding mathematics.
P2	Range of conflicting requirements	Balancing image quality, model complexity, embedding dimensionality, and computational limitations.
P3	Depth of analysis required	Needed multiple evaluation approaches—visualization, similarity scoring, and model comparison.
P4	Familiarity of issues	Involved multiple unfamiliar domains such as molecular data integration, dataset biases, and skin tone variability.
P5	Extent of applicable codes	No established standard for combining dermatology with phytochemical VLMs.
P6	Interdependence	Involves interconnected systems—vision models, text models, embeddings, and retrieval functions.

7.2 Complex Engineering Activities(CEA)

TABLE III. A SAMPLE COMPLEX ENGINEERING PROBLEM ACTIVITIES TABLE

Attributes		Addressing the Complex Engineering Activities(A)
A1	Range of resources	Requires human collaboration, GPU hardware, datasets, modern AI tools.
A2	Level of interactions	Collaboration among team members for integration, debugging, and testing.
A3	Consequences to society/environment	Potential for natural skincare solutions and reduced environmental harm.
A4	Familiarity	Involves learning new tools, datasets and biochemical knowledge domains.

Chapter 8 Conclusions

8.1 Summary

In this project, a multimodal vision-language interface between dermatological images and plant-derived bioactive products was created. The incorporation of DermNet, SkinCon and HERB 2.0 datasets together with models like CLIP, SigLIP and BLIP2 resulted in meaningful embeddings which allowed the system to perform cross-modal retrieval. The project shows that AI can assist in the initial research on natural skincare solutions by revealing possible compound-condition associations.

8.2 Limitations

- Image datasets contain noise, low-quality samples, and class imbalance.
- HERB 2.0 descriptions vary in detail and quality.
- Multimodal models were not fine-tuned on domain-specific medical data.
- Retrieval accuracy is qualitative rather than clinically validated.
- Lacks large-scale benchmarking and expert dermatological evaluation.

8.3 Future Improvement

- Fine-tune VLMs using dermatology-specific data.
- Include clinical metadata and patient history for deeper reasoning.
- Expand the herbal compound database.

- Train a supervised model for disease-to-herb prediction.
- Add interpretability tools like Grad-CAM for medical explanation.
- Develop a web-based interface for real-world usability.

References

[1] K. Michalak, “Phytochemicals as skin-protective agents against oxidative stress and inflammation,” *J. Phytomedicine,* 2022.
[2] R. Anbualakan et al., “Role of plant-derived flavonoids and carotenoids in skin photoprotection,” *Dermatol. Res. Pract.,* 2022.
[3] Y. Zhao, “Medicinal Plant Extracts Targeting UV-Induced Skin Damage,” *J. Nutr. Biochem.*, 2025.
[4] M. Tomas, “Plant-based phytochemicals for skin care,” *Immunity & Ageing*, vol. 25, no. 4, 2025.
[5] I. Dammak et al., “Phytochemicals in skin health,” *Int. J. Mol. Sci.,* 2025.
[6] R. Fang et al., “HERB 2.0: An updated database integrating evidence for traditional medicine,” *Nucleic Acids Res.,* 2024.
[7] J. Li et al., “BLIP-2: Bootstrapping Language–Image Pre-training,” arXiv:2301.12597, 2023.
[8] C. Pan et al., “A multimodal vision foundation model for clinical dermatology,” *Nature Medicine*, 2025.

[9] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, 2021.

[10] A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

[11] J. Li et al., “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” arXiv:2301.12597, 2023.

[12] X. Zhai et al., “Sigmoid Loss for Language Image Pre-training,” ICCV, 2023.

[13] C. D. Manning et al., Introduction to Information Retrieval, Cambridge Univ. Press, 2008.

[14] K. Järvelin and J. Kekäläinen, “Cumulative Gain-based Evaluation of IR Techniques,” ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, 2002.

[15] A. Vaswani et al., “Attention Is All You Need,” NeurIPS, 2017.