Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Evaluating The Embedding Space of Foundation Models for Dermatological Images to Guide Backbone Selection for A Fine-Tuning Pipeline

Thi Anh Thu Pham [1], Khoi Nguyen Gia[1], Tran Hieu [1], Tran Anh Duc [1], Thi Thanh Tu Bui[1]
[1]University of Foreign Languages - Information Technology, HCMC, Vietnam

Abstract **- In recent years, computer vision models have advanced rapidly and have been increasingly applied in dermatology. However, in practice, many skin diseases present with very similar visual manifestations, making image representation in vector form not always accurately reflect the true nature of the disease. Therefore, this study focuses on comparing two approaches: general-purpose models designed for multiple tasks and models specifically trained for medical and dermatological domains.Experiments were conducted on 1,300 dermatological images covering 13 different infectious diseases, with a balanced number of images across disease categories. The models were evaluated based on their ability to retrieve images with similar characteristics, while also considering computational cost.The results show that in terms of pure representation quality, an ultra-large-scale general-purpose model achieved the highest performance. However, dermatology-specific models provided an optimal balance between performance and computational cost for the infectious disease group, producing stable and clinically appropriate image representations to serve as a foundation for subsequent clinical research.**

Keywords - Dermatology Embedding; Embedding Evaluation; Foundation Models; Backbone Selection; Skin Retrieval

I. INTRODUCTION

Dermatological images are one of the most important sources of information in the examination and treatment of skin diseases. In clinical practice, physicians often make initial assessments based on direct visual inspection of lesion morphology, including location, size, color, and changes in skin surface characteristics. Beyond their diagnostic support role, dermatological images also help monitor disease progression over time—from the onset stage to treatment stabilization (Mehta et al., 2025)—thereby supporting the evaluation of intervention effectiveness and long-term patient management. For this reason, effectively leveraging information from dermatological images is of significant importance for both clinical practice and scientific research.

Alongside the rapid development of image processing and machine learning methods in recent years, numerous automated tools have been proposed to support medical image analysis. A common approach is to convert images into vector representations to facilitate comparison and recognition of similar patterns. In deep learning for dermatological image analysis, Convolutional Neural Networks (CNNs), transformer models, and embedding vectors are widely used to extract image features for classification and similar-image retrieval tasks, and have demonstrated effectiveness in many real-world applications (Mehta et al., 2025). However, when directly applied to dermatological images, these general-purpose models gradually reveal notable limitations, such as performance degradation when handling out-of-distribution data, limited ability to capture diversity in skin tones or rare diseases, and excessive dependence on standardized datasets (Daneshjou et al., 2022).

Unlike many conventional computer vision tasks, the differences between skin diseases often do not lie in easily recognizable features. In practice, the manifestations of many skin conditions can appear highly similar, especially in early stages or when lesions occur at different anatomical locations, resulting in high inter-class similarity and substantial intra-class variation—posing significant challenges for traditional deep learning networks in accurately separating disease categories (Mehta et al., 2025). The distinctions that clinicians focus on are often subtle, such as whether lesion borders are well-defined, whether the skin surface is scaly or moist, or whether lesion coloration is evenly distributed or patchy. These are micro-level features expressed at fine image detail. For models trained primarily on everyday natural images, such subtle characteristics are not the salient features they are designed to prioritize, thereby reducing their ability to accurately differentiate complex dermatological lesions (Wen et al., 2024).

To address these challenges, recent research has shifted toward developing models trained directly on medical data, including dermatological images (Shrestha et al., 2023). These models are expected to learn features that are more aligned with clinical contexts rather than relying solely on superficial visual similarity. Preliminary findings suggest that domain-specific models are better able to cluster images according to underlying pathological characteristics and reduce confusion between diseases with visually similar presentations (Yan, Yu, & Primiero, 2025). However, most existing studies still focus primarily on final disease classification performance, while the role of image representation as a foundational component for subsequent research directions—such as model fine-tuning or the development of decision-support systems—has not yet been thoroughly and systematically examined (Q. Chen et al., 2024)

Against this backdrop, the present study concentrates on examining and comparing different image representation approaches applied to dermatological images, with the aim of evaluating their suitability in reflecting clinically meaningful features. Rather than focusing solely on classification accuracy, the study emphasizes the ability to retrieve images with similar

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

characteristics and the way images are organized within the representation space. Through this analysis, the research seeks to identify models with the potential to serve as foundational backbones for future dermatology-related research and applications (Yan, Hu, & Jiang, 2025).

II. RESEARCH METHODOLOGY

A. RESEARCH OBJECTIVES

This study was conducted with the objective of rigorously evaluating the practical suitability of existing embedding representation models when directly applied to dermatological images from infectious disease categories, in a setting where additional labeled training data are minimal or nearly unavailable. Rather than following a traditional classification-based approach—which requires fully annotated datasets and computationally expensive training procedures—the study raises a question that is closer to real-world deployment: whether embedding models, trained on different data sources and knowledge domains, are capable of naturally organizing dermatological images such that images of the same pathology are positioned close to one another in the representation space. If the embedding space fails to reflect these pathological relationships, any downstream applications—from similar-image retrieval to diagnostic support—will be fundamentally constrained at the representation level.

This requirement is particularly critical for content-based image retrieval (CBIR) systems—an approach that has gained increasing attention in medicine—where clinicians and researchers may compare a new case with previously documented similar cases. Fundamentally, these systems do not rely on predefined disease labels but instead operate by converting each image into a numerical representation and computing distances between representations within a shared space. When a new image is introduced, the system retrieves the most similar embeddings, thereby providing reference cases with comparable characteristics.

In the medical domain, several studies have shown that this approach can remain effective even without retraining the original model, provided that the embedding representations are sufficiently expressive to capture clinically relevant image features (Khun Jush et al., 2023). This has substantial practical significance, as labeled medical data are not only scarce but also require high levels of expertise and considerable cost to curate (Lee et al., 2023). In this context, leveraging pretrained models as ready-made feature extractors becomes a more feasible and pragmatic alternative than developing models from scratch (Raghu et al., 2019). Accordingly, this study employs embedding models in their original pretrained state, treating them as fixed feature extractors without further training or parameter fine-tuning (Esteva et al., 2017). This design choice reflects common real-world deployment scenarios, in which pretrained models are often used for rapid data assessment, prototype retrieval system development, or as a preprocessing step before deciding whether deeper domain-specific training is warranted (Rashad et al., 2023). Prior medical research has likewise demonstrated that pretrained embeddings can be directly applied for meaningful image feature extraction, particularly in retrieval and similarity analysis tasks (Khun Jush et al., 2023).

Based on this rationale, the study compares models from three distinct categories—general-purpose models, biomedical models, and dermatology-specific models—in order to clarify differences in how they structure the embedding space when applied to the same dataset of infectious dermatological diseases. In practice, differences in the training data domain have repeatedly been identified as a key factor influencing the transferability of embeddings to new tasks. Therefore, the evaluation does not merely aim to determine which model performs better, but rather addresses a more strategic question: under data-constrained conditions, which type of model should be prioritized as the foundational backbone for dermatological applications. In this sense, the study serves as a necessary preliminary evaluation, providing a basis for selecting suitable embedding models prior to more advanced development stages, including the construction of image retrieval systems, similarity analysis of pathological manifestations, or fine-tuning for future dermatological diagnostic tasks.

B. DATASET

To ensure objectivity and reproducibility, the dataset used in this study consists of 1,300 dermatological images carefully collected and curated from reputable public medical image repositories specializing in clinical photography, including DermNet (DermNet, n.d.), Fitzpatrick17k (Groh et al., 2021), and SD-198 (Nguyen, n.d.). A key criterion during dataset aggregation was strict consistency in image modality: all selected samples are clinical macro images of the skin surface captured using standard optical devices (such as smartphones or digital cameras), with complete exclusion of dermoscopy and histopathology images. This uniformity minimizes technical noise arising from differences in acquisition equipment, thereby compelling the embedding models to focus on true pathological morphological features under conditions that closely resemble physicians' direct visual inspection.

The dataset is evenly distributed across 13 distinct diseases, all belonging to the infectious disease category. Each disease includes exactly 100 images to ensure class balance, thereby reducing the risk of evaluation bias caused by class size disparities. Maintaining a balanced dataset structure ensures that comparative experiments genuinely reflect each model's ability to organize the embedding space, rather than being influenced by unequal data volumes among disease classes.

The selected diseases represent common infectious dermatological conditions encountered in clinical practice, spanning multiple etiological categories such as viral, bacterial, fungal, and parasitic infections. Specifically, the list includes: Chickenpox, Folliculitis, Head Lice, Herpes, Herpes Zoster, Infectious (unspecified infections), Infestations/Bites, Monkeypox, Nail Fungus, Ringworm, Sarampion (Measles), Tinea Capitis, and Warts. This selection does not aim to exhaustively cover the entire dermatological spectrum but instead focuses on diseases with clear external manifestations that are typically preliminarily diagnosed and monitored visually in routine clinical settings.

Restricting the scope to infectious diseases is motivated not only by their prevalence in clinical practice but also by the distinctive challenges they pose for artificial intelligence systems. In reality, many diseases within this group exhibit highly similar external appearances, particularly in early stages or when lesions occur on different anatomical regions. Clinical

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

signs such as erythema, vesicles, scaling, plaque-like lesions, or small papules may be shared across multiple conditions. As a result, diagnostic boundaries become blurred if a model relies solely on low-level visual features without capturing deeper medical semantics.

This morphological similarity makes the infectious disease group a stringent testbed for evaluating embedding model quality. A poorly structured embedding space—one that depends primarily on color or global shape—will easily miscluster visually similar but pathologically distinct diseases. In contrast, a well-structured embedding space should demonstrate the ability to "pull" images of the same pathology closer together despite variations in viewing angle, lighting conditions, or skin tone, while simultaneously "pushing" apart images of different diseases even when they appear visually similar. In this context, the dataset serves not merely as input data but as a high-realism evaluation environment that accurately reflects the semantic barriers dermatological AI systems will encounter in real-world hospital deployment.

C. EMBEDDING MODELS EVALUATED

The embedding models selected for this study are categorized into three main groups based on the domain of data on which they were originally trained: general-purpose models, general medical models, and dermatology-specific models. This categorization is not merely technical; it reflects a core assumption of the study: that the degree of alignment between the training domain and the application domain directly influences how a model organizes and represents information within the embedding space.

The general-purpose model group consists of models trained on large and diverse datasets, primarily composed of everyday images and online visual content. The strength of this group lies in its ability to learn broadly applicable visual features, enabling stable performance across various image types. However, because these models were not trained with medical objectives in mind, they tend to emphasize easily recognizable characteristics such as overall shape, prominent colors, or general image composition. In dermatology—where diagnostically meaningful differences are often subtle—the critical question is whether such general visual features are sufficient to differentiate and meaningfully associate pathological manifestations.

The general medical model group aims to narrow this gap by being trained on multidisciplinary medical imaging datasets. These models are expected to be more sensitive to features commonly present in medical images, such as tissue structures, density variations, or lesion morphology. However, because their training data span a wide range of imaging modalities—including X-ray, CT, MRI, and endoscopy—their suitability for dermatological images, which have distinct characteristics in terms of color, surface texture, and photographic context, must be carefully evaluated.

Finally, the dermatology-specific model group is trained directly on datasets of skin lesions, with the objective of capturing morphological features that are clinically meaningful in dermatological diagnosis. These models are expected to organize the embedding space more strongly around pathological relationships rather than relying solely on superficial visual similarity. Including this group in the comparison does not presume absolute superiority; instead, it serves to examine whether domain specialization in training truly translates into advantages within the embedding space—particularly in a setting without retraining or parameter fine-tuning.

By positioning these three model groups side by side, the study enables a direct analysis of the relationship between training data domain and the quality of the resulting embedding space when applied to infectious dermatological images. Rather than focusing solely on final performance metrics, this approach helps clarify why a given model performs better or worse in a specific context. Detailed information about the models used—including version, embedding dimensionality, and inference-time characteristics—is presented in the results section to ensure transparent and systematic comparison and analysis.

D. SIMILARITY METRIC SELECTION

In embedding-based image retrieval systems, measuring how "similar" two images are is not as simple as computing the geometric distance between two points in space. More importantly, the chosen distance must meaningfully reflect semantic similarity. An inappropriate metric may cause the system to return images that are "mathematically close" yet "semantically distant," particularly in complex domains such as dermatological imaging, where distinct diseases may exhibit highly similar visual morphology.

Euclidean distance has traditionally been used in retrieval systems due to its intuitive interpretation: the closer two vectors are, the more similar the corresponding images are assumed to be. This assumption holds if both vector magnitude and absolute position carry semantic meaning, as in classical feature engineering systems where feature strength is directly tied to image content. However, this assumption no longer holds for modern embedding models. In contrastive or multimodal models such as CLIP (Yan et al., 2026), the training objective is not to make similar images close in every coordinate dimension, but rather to align them in direction within the embedding space. In other words, the model encodes semantic similarity primarily through vector orientation rather than vector magnitude. Consequently, embeddings are often normalized and distributed on a hypersphere during training.

Under this paradigm, continuing to use Euclidean distance introduces a conceptual mismatch: it measures a quantity that the model was not explicitly optimized to preserve. Two vectors may be highly aligned in direction—indicating the same dermatological pathology—but differ slightly in magnitude, resulting in a substantial Euclidean distance despite negligible medical or semantic difference. Cosine similarity directly addresses this issue by discarding magnitude information and preserving only directional alignment. When cosine similarity is applied, the system effectively answers the question:

"Does the model interpret these two images as representing the same semantic content?"

rather than:

"How many geometric units apart are these two vectors?"

This distinction is especially critical for dermatological images. The same pathology may be photographed across different anatomical sites, lighting conditions, or skin tones,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

leading to variations in embedding magnitude without altering the lesion's intrinsic clinical nature. Cosine similarity enables the system to focus on stable, clinically meaningful information rather than irrelevant magnitude fluctuations.

An additional advantage of cosine similarity emerges in multimodal embedding spaces, where image and text representations coexist. In such cases, image and text vectors may have inherently different magnitudes due to modality-specific encoding processes, yet still "point" toward the same pathological concept. Cosine similarity allows consistent comparison across modalities, whereas Euclidean distance is highly sensitive to differences in vector norms, amplifying the modality gap.

Mathematically, let u and v be two embedding vectors in $\mathbb{R}^n$ representing an image–image or text–image pair. Euclidean distance measures the absolute geometric distance between them:

$$d_{Euclid}(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$$

While intuitive, this metric is directly influenced by vector magnitude.

Cosine similarity, in contrast, measures the angle between vectors:

$$Cosine(u,v) = \frac{u \cdot v}{|u|_2|v|_2} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

The numerator captures directional alignment (dot product), while the denominator normalizes by L2 norms, eliminating magnitude effects. The cosine similarity value lies within [-1, 1]. When two embeddings share the same pathological semantics, their vectors tend to align ($\theta \rightarrow 0°$), and the cosine value approaches 1.

When vectors are L2-normalized ($\|u\|_2 = \|v\|_2 = 1$), Euclidean distance and cosine similarity become mathematically related and monotonically connected (Bishop, 2006). However, in high-dimensional, non-uniform embedding spaces—especially multimodal ones—cosine similarity consistently demonstrates superior stability, as it neutralizes discrepancies in vector norms that frequently arise across modalities.

To empirically validate the appropriateness of cosine similarity in this specific dataset, a rapid evaluation experiment was conducted on the same 1,300 images under a cross-modal (text-to-image) retrieval configuration. The comparative results revealed substantial performance degradation when Euclidean distance replaced cosine similarity. For instance, with OpenAI CLIP ViT-Base/32, Top-1 retrieval accuracy using cosine similarity reached 0.2938, but dropped to 0.1192 when Euclidean distance was used (a reduction of over 59%). A similar trend was observed for OpenAI CLIP ViT-Large/14, where Mean Top-1 decreased from 0.4084 (cosine) to 0.0746 (Euclidean).

This severe degradation clearly illustrates the mathematical limitation of Euclidean distance in multimodal embedding spaces. Image and text encoders process fundamentally different data streams, producing vectors with significantly different magnitudes and creating a modality gap. Euclidean distance inadvertently amplifies this discrepancy, transforming it into retrieval noise. By contrast, foundation models such as CLIP are optimized using contrastive loss to align the orientation of corresponding image and text embeddings. Therefore, cosine similarity—by neutralizing magnitude and focusing exclusively on vector angles—is fully aligned with the model's semantic learning mechanism.

Based on both architectural theory and empirical evidence, cosine similarity is formally selected as the standard similarity metric for the entire evaluation pipeline.

E. IMILARITY METRIC SELECTION

After extracting embeddings for each image—and, where applicable, for textual prompts in the format "a photo of …"—the next step is to determine the similarity metric used to compare these vectors. This decision is not merely technical; it reflects a geometric and semantic interpretation of the embedding space, as well as the way embedding models are designed and trained.

In embedding-based image retrieval, the two most commonly used similarity measures are Euclidean distance and cosine similarity. Euclidean distance computes the straight-line ("as-the-crow-flies") geometric distance between two vectors in vector space (T. Chen et al., 2020), whereas cosine similarity measures the angle between two vectors—i.e., their directional alignment—independent of vector magnitude (Qian et al., 2004).

A key point emphasized in embedding literature is that when embeddings are normalized—a common practice in deep learning models and retrieval systems—cosine similarity more directly reflects semantic similarity than Euclidean distance (You, 2025). Specifically, cosine similarity considers only the angle between vectors and is unaffected by magnitude variations, which may arise from factors such as image quality, illumination conditions, or technical properties of the input data. This makes cosine similarity more stable in real-world environments, where vectors encoding the same semantic meaning may differ in magnitude but should still be considered similar (T. Chen et al., 2020).

Modern embedding models—particularly those trained using contrastive learning objectives, such as CLIP and its variants—are optimized to align vectors of semantically related content in terms of direction rather than absolute magnitude. Therefore, cosine similarity is better aligned with the operational principles of these models, as it directly corresponds to the semantic structure learned during training.

Several technical analyses further indicate that in high-dimensional embedding spaces (often ranging from hundreds to thousands of dimensions), Euclidean distance becomes less stable (T. Chen et al., 2020) and more sensitive to vector magnitude, whereas cosine similarity focuses on directional relationships and is thus less affected by technical variations in the data (Radford et al., 2021). This observation is also supported in experimental reports on vision–language retrieval systems (Radford et al., 2021). From a practical standpoint, cosine similarity is particularly advantageous in multimodal settings, where image and text embeddings are mapped into a shared vector space, enabling direct comparison across modalities without dependence on vector norms.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

In summary, cosine similarity was selected in this study for the following reasons:

- Alignment with embedding learning mechanisms: Semantic similarity is primarily encoded in vector orientation rather than magnitude, especially in contrastive learning models.
- Robustness to real-world magnitude variations: It remains stable under fluctuations caused by image quality, image size, or illumination differences—common characteristics in dermatological image datasets.
- Compatibility with multimodal embeddings: It enables consistent comparison between image and text representations within a shared embedding space.

Accordingly, cosine similarity is well suited to the study's objective of evaluating the semantic organization capability of embedding spaces in dermatology image retrieval.

Beyond the choice of similarity metric, the retrieval evaluation protocol itself plays a crucial role. In this study, retrieval performance is measured using the Top-K Accuracy metric. In the medical context, Top-K reflects the percentage of queries for which at least one image from the same disease class as the query appears within the top K returned results.

Combining Top-K evaluation with cosine similarity provides a comprehensive perspective on embedding space structure. While cosine similarity ensures that ranking is not distorted by vector magnitude differences, Top-K—particularly at K = 5, 10, and 20—assesses the model's ability to form meaningful local semantic clusters. Even if the model fails at Top-1 due to extreme morphological similarity among infectious diseases, the presence of a correct match within Top-5 or Top-10 indicates that the embedding space successfully pulls semantically related lesions closer together.

This evaluation strategy closely aligns with clinical practice, where physicians often review a group of similar cases before reaching a final diagnosis, rather than relying on a single nearest match.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. GENERAL MODEL GROUP (GENERAL MODELS

This group of models includes variants of CLIP (Contrastive Language–Image Pretraining) trained on large-scale Internet datasets, whose original objective was not specifically targeted at the medical domain or dermatology. Therefore, their performance in the context of pathological images directly reflects the models' visual generalization ability rather than their domain suitability.

TABLE 1 Performance of General-Purpose Models

STT	Model Name	Model Version	Embedding Dim	Total Images	Average Search Time	Average Embedding Time	Average Total Time	Mean Top 1	Mean Top 2	Mean Top 5	Mean Top 10	Mean Top 20	Average Accuracy	Model Type
1	AltCLIP	BAAI/AltCLIP	768	1300	0.000339	0.059907	0.060245	0.363846	0.491538	0.676154	0.880769	1	0.363846	General Model
2	ConvNeXt-CLIP	convnext_base_w-laion2b	640	1300	0.000218	0.01666	0.016878	0.368462	0.507692	0.715385	0.896154	1	0.368462	General Model
3	MetaCLIP	H14-FullCC2.5B	1024	1300	0.000319	0.146832	0.147151	0.513846	0.683846	0.823077	0.983077	1	0.513846	General Model
4	MobileCLIP	S2	512	1300	0.000259	0.02941	0.029669	0.386154	0.516923	0.684615	0.860769	1	0.386154	General Model
5	OpenAI-CLIP	ViT-Base/32	512	1300	0.000213	0.015455	0.015668	0.293846	0.410769	0.621538	0.817692	1	0.293846	General Model
6	OpenAI-CLIP	ViT-Large/14	768	1300	0.000236	0.062079	0.062315	0.408462	0.512308	0.719231	0.906154	1	0.408462	General Model
7	Google-SigLIP	so400m-patch14-384	1152	1300	0.000344	0.268847	0.269191	0.511538	0.676154	0.823077	0.946154	1	0.511538	General Model

The experimental results in Table 1 reveal a consistent trend: the larger the training dataset and model size, the better the retrieval performance, even when the model is not fine-tuned for the medical domain. This indicates that sufficiently large general-purpose models are still capable of learning fundamental visual representations that can be reused for specialized tasks such as dermatology.

Specifically, MetaCLIP (Xu et al., 2023) and SigLIP (Zhai et al., 2023) achieve the highest performance within this group, with Top-1 accuracies of 51.38% and 51.15%, respectively. For the Top-5 metric, both models exceed 82%, indicating that when more retrieval results are allowed, the probability of retrieving an image with the same pathology increases substantially. These findings suggest that models trained on large-scale Internet data are capable of effectively learning basic visual features such as color, patterns, texture, and overall shape—factors that play a critical role in distinguishing skin lesions, even when multiple diseases exhibit highly similar morphological characteristics.

In contrast, OpenAI CLIP ViT-B/32 (OpenAI, 2021a) shows the lowest performance in the group, with a Top-1 accuracy of only 29.38%. The performance gap of more than 20% compared to MetaCLIP clearly highlights an important limitation: a smaller architecture and more limited pre-training data are insufficient to produce a highly discriminative embedding space for complex disease categories. This limitation is particularly evident in infectious dermatology tasks, where morphological boundaries between diseases are often subtle and poorly defined.

However, high performance is not the only factor to consider. The embedding extraction time results demonstrate a clear trade-off between accuracy and computational cost. Although SigLIP achieves strong performance, its processing time reaches 0.269 seconds per image—approximately 17 times slower than OpenAI CLIP (0.015 seconds per image) (OpenAI, 2021a). In real-world deployment scenarios, especially large-scale image retrieval systems or clinical applications requiring rapid responses, this computational cost may become a significant constraint.

In summary, the results of the general-purpose model group indicate that model scale and training data size play a decisive role in embedding quality, even across distinct domains such as dermatology. Nevertheless, higher performance comes with increased inference cost, raising practical concerns regarding deployment feasibility. This motivates further comparison with general medical models and dermatology-specific models, which may achieve a more favorable balance among accuracy, domain suitability, and computational efficiency

B. GENERAL MEDICAL MODEL GROUP (MEDICAL MODELS)

This group of models includes variants of CLIP (Contrastive Language–Image Pretraining) trained on large-scale Internet datasets, whose original objective was not specifically targeted at the medical domain or dermatology. Therefore, their performance in the context of pathological images directly reflects the models' visual generalization ability rather than their domain suitability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

TABLE 2. Performance of Medical Models

STT	Model Name	Model Version	Embedding Dim	Total Images	Average Search Time	Average Embedding Time	Average Total Time	Mean Top 1	Mean Top 2	Mean Top 5	Mean Top 10	Mean Top 20	Average Accuracy	Model Type
8	BiomedCLIP	PubMedBERT_256-ViT_Base_16	512	1300	0.000222	0.011186	0.011408	0.253077	0.43	0.63	0.884615	1	0.253077	Medical Model
9	ClipMD	Idan0405/ClipMD	512	1300	0.000246	0.01753	0.017776	0.179231	0.299231	0.576923	0.883077	1	0.179231	Medical Model
10	MedImageInsights	2024.09.27	1024	1300	0.000321	0.232241	0.232562	0.38	0.558462	0.783077	0.956923	1	0.38	Medical Model
11	MedSigLIP	448	1152	1300	0.000399	0.312309	0.312708	0.464615	0.592308	0.722308	0.948462	1	0.464615	Medical Model
12	PLIP	vinid/plip	512	1300	0.000216	0.015066	0.015282	0.163846	0.266154	0.524615	0.866154	1	0.163846	Medical Model
13	RCLIP	kaveh/rclip	512	1300	0.00026	0.066331	0.06659	0.087692	0.16	0.420769	0.784615	1	0.087692	Medical Model

Contrary to expectations, most medical models—such as PLIP (vinid, 2023) and RCLIP (Shahhosseini, 2023)—perform poorly on dermatology images. The primary cause is domain misalignment: X-ray and MRI data differ substantially from color dermatology images.

Table 2 highlights a paradox: models labeled as "medical" do not necessarily perform effectively on dermatology images. The performance degradation caused by domain shift is evident. Models such as PLIP (Pathology) (vinid, 2023) and RCLIP (Radiology) (Shahhosseini, 2023) achieve very low results, with Top-1 accuracies of 16.38% and 8.77%, respectively.

The core reason lies in the fundamental differences in data characteristics. PLIP was trained on pathology images (microscopic tissue images), while RCLIP was trained on radiology images (X-ray/CT scans—grayscale images of internal organs). The embedding spaces of these models have been optimized for features that do not exist in RGB skin surface images, leading to a phenomenon known as negative transfer.

An exception is MedSigLIP (Sellergren et al., 2025), which maintains competitive performance with a Top-1 accuracy of 46.46%. This can be attributed to the strong SigLIP backbone architecture and a more diverse medical training corpus (including clinical images), preventing the model from being "blind" to dermatological features. However, it is also the slowest model in the entire experiment, with a processing time of 0.312 seconds per image.

C. *DERMATOLOGY-SPECIFIC MODEL GROUP (SKIN DISEASE MODELS)*

This group consists of models that are fine-tuned or retrained directly on dermatology datasets.

TABLE 3. Performance of Dermatology-Specific Models

STT	Model Name	Model Version	Embedding Dim	Total Images	Average Search Time	Average Embedding Time	Average Total Time	Mean Top 1	Mean Top 2	Mean Top 5	Mean Top 10	Mean Top 20	Average Accuracy	Model Type
14	DermLIP	ViT-B-16	512	1300	0.000228	0.012957	0.013185	0.493077	0.630769	0.790769	0.947692	1	0.493077	Skin Disease
15	WhylesionCLIP	ViT-L-14 (HF-Hub)	768	1300	0.000255	0.079861	0.080116	0.319231	0.443846	0.700769	0.926154	1	0.319231	Skin Disease

The dermatology-specific model group demonstrates the greatest potential in balancing accuracy and system efficiency. **DermLIP (Yan, Yu, Primiero, & Vico-Alonso, 2025) – The Optimal Trade-off:**DermLIP (ViT-B/16) achieves a Top-1 accuracy of **49.31%**. Although this is approximately 2.07% lower than the leading model, MetaCLIP (51.38%), the difference on a dataset of 1,300 images is relatively small and may not be statistically significant. It is important to note that MetaCLIP H14 is a massive architecture, whereas DermLIP relies only on a ViT-Base backbone. The fact that a comparatively smaller model can closely match a super-large model clearly demonstrates the power of training on domain-specific data. Notably, DermLIP's processing time is only **0.013 seconds per image**, making it approximately **11 times faster than MetaCLIP** and **24 times faster than** MedSigLIP (Sellergren et al., 2025). **Significance of Top-K Metrics:** When expanding the retrieval scope to Top-5 and Top-10, DermLIP achieves accuracies of **79.07%** and **94.76%**, respectively. The **Mean Top-20 score reaches 1.0 (100%)**, indicating that although the model may occasionally misrank the correct label at the first position (due to strong inter-class similarity), the correct label consistently lies within a very close neighborhood in the vector space. This confirms that DermLIP's semantic structure is highly coherent and stable. **Failure of WhylesionCLIP (Yang et al., 2024):** Despite being dermatology-specific, WhylesionCLIP achieves only **31.92% Top-1 accuracy**. This suggests that simply using dermatology data is insufficient; training strategy and data curation play a crucial role. A larger embedding size (768 dimensions compared to DermLIP's 512) does not compensate for shortcomings in learning core discriminative features.

D. *EVALUATION OF THE ROLE OF THE COSINE SIMILARITY ALGORITHM*

Overall, for certain models such as DermLIP (Yan, Yu, Primiero, & Vico-Alonso, 2025) and MetaCLIP (Xu et al., 2023), the performance difference between the two similarity metrics is negligible. However, a pivotal discrepancy emerges within the OpenAI-CLIP model group. Experimental results reveal a severe accuracy drop when replacing Cosine similarity with Euclidean distance. For OpenAI CLIP ViT-B/32 (OpenAI, 2021a), the Top-1 accuracy sharply declines from 29.38% (Cosine) to only 11.92% (Euclidean). The negative impact of Euclidean distance becomes even more pronounced in OpenAI CLIP ViT-L/14 (OpenAI, 2021b). Specifically, the Top-1 accuracy drops dramatically from 40.84% (Cosine) to 7.46% (Euclidean). This reduction of more than 80% clearly demonstrates the high sensitivity of the embedding space to the choice of distance metric, highlighting the critical role of similarity computation in retrieval performance.

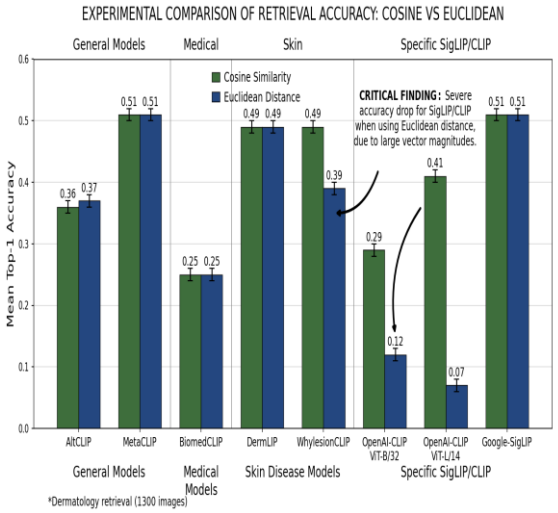

Fig 1. Experimental comparison of retrieval accuracy between Cosine similarity and Euclidean distance

Within the scope of this study, the reported Top-K metrics are not intended solely for comparing raw performance; they also demonstrate the decisive role of selecting an appropriate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

distance metric for ranking tasks. The highest-performing models—such as MetaCLIP (Xu et al., 2023) and DermLIP (Yan, Yu, Primiero, & Vico-Alonso, 2025)—are contrastive learning models optimized in terms of angular relationships (vector direction).

When Euclidean distance is used, variations in brightness, contrast, or shooting angle in clinical dermatology images alter the vector norm (magnitude), unintentionally "pushing" images of the same pathology farther apart. As a result, correct matches are downgraded in ranking, leading to a severe decline in Top-K performance (as evidenced by the more than 80% drop observed in OpenAI CLIP ViT-L/14 (OpenAI, 2021b)).

In contrast, Cosine Similarity focuses solely on the vector angle and thus acts as a naturally noise-robust mechanism. This preserves the ranking of semantically similar pathological images, explaining why the system maintains absolute stability from Top-1 to Top-20 (e.g., the Mean Top-20 score of 1.0 achieved by DermLIP).

This finding is particularly meaningful in the dermatology context. In real-world datasets, input images exhibit substantial variation in lighting conditions, camera angles, and magnification levels (macro vs. full-field). These physical differences inevitably affect the magnitude of feature vectors. However, by employing Cosine Similarity—which considers only vector direction while disregarding magnitude—the retrieval system implicitly establishes a natural noise-resistant mechanism.

The successful experimental results confirm that using Cosine Similarity is not merely a technically appropriate decision; it is also a fully compatible configuration for maximizing the semantic knowledge learned by foundation models, while neutralizing undesirable variations inherent in clinical medical images

E. CONCLUSIONS DRAWN FROM THE DATA

From the quantitative analyses above, the study draws three main conclusions to guide backbone selection:

- DermLIP (Yan, Yu, Primiero, & Vico-Alonso, 2025) provides the most practical balance for real-world applications due to its superior accuracy/cost ratio, even though it does not achieve the absolute highest accuracy.
- MetaCLIP (Xu et al., 2023) and SigLIP (Zhai et al., 2023) are well suited as teacher models or for offline tasks requiring extremely high accuracy without strict time constraints.

Models trained on mismatched medical domains—such as PLIP (vinid, 2023) and RCLIP (Shahhosseini, 2023)—should be excluded from dermatology tasks, as they provide no added value and, in fact, reduce accuracy compared to general-purpose models.

IV. DISCUSSION

This study was designed to evaluate the quality of the embedding space produced by different models when applied to infectious dermatology images, rather than to directly optimize classification accuracy. Therefore, the experimental results should be interpreted from the perspective of morphological semantic representation and domain suitability, rather than through a purely numerical comparison of accuracy values alone.

A. IMPACT OF TRAINING DATA DOMAIN ON EMBEDDING QUALITY

The experimental results reveal a clear distinction among the three model groups identified at the outset: general-purpose models, general medical models, and dermatology-specific models. Overall, models trained on data more closely aligned with the dermatology domain tend to produce embedding spaces that are more stable, semantically structured, and better suited for dermatological image retrieval tasks.

Within the general-purpose model group, large-scale backbones such as MetaCLIP and Google-SigLIP achieve relatively strong performance compared to traditional CLIP variants. This suggests that large-scale training data and visual diversity can partially compensate for the lack of domain-specific knowledge. However, these models still tend to rely heavily on generic visual features such as color, contrast, or overall shape. Such features are insufficiently fine-grained to effectively distinguish infectious skin diseases with similar morphological presentations, where differences often lie in subtle microstructural patterns and lesion distribution.

In contrast, general medical models do not demonstrate the expected advantage. Although trained on biomedical data, these models typically cover multiple specialties—such as X-ray, CT, MRI, and histopathology—resulting in dermatological representations that are not specifically optimized. This limitation is reflected in the decline of Top-K accuracy metrics compared to several strong general-purpose models, indicating that broad medical knowledge alone is insufficient to address the highly specialized characteristics of dermatological tasks.

Most notably, the dermatology-specific group stands out. DermLIP demonstrates competitive and stable performance across most evaluation metrics. Although its Top-1 accuracy does not surpass that of the very large-scale MetaCLIP architecture, the result is practically significant when considering the substantial differences in parameter scale and computational cost. Training directly on skin lesion images enables the model to learn clinically meaningful micro-morphological features, such as lesion boundaries, surface structures, scaling patterns, infiltration levels, and lesion distribution patterns.

Taken together, these findings reinforce the central hypothesis of the study: the alignment between training data domain and target application is a decisive factor in determining embedding space quality for dermatological tasks. While model scale and large datasets can enhance performance, they cannot fully substitute for the advantages of domain-specific training.

B. IMPACT OF RESTRICTING THE EVALUATION TO INFECTIOUS DISEASES

An important point that must be emphasized is that the dataset used in this study includes only infectious dermatological conditions. This category is characterized by a high degree of morphological overlap and commonly appears in similar anatomical regions such as the scalp, trunk, extremities, or intertriginous areas. As a result, the task becomes considerably more challenging than dermatological

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

classification settings that include inflammatory diseases, benign tumors, malignant tumors, and normal skin.

In this context, the embedding space is required not only to differentiate across broad disease categories but also to separate conditions with highly similar visual characteristics in terms of color, texture, and lesion morphology. Consequently, moderate Top-1 accuracy values should not be interpreted as a limitation of the models. Rather, they should be understood as an indicator of the intrinsic difficulty of the task under realistic clinical conditions.

Conversely, the fact that many models achieve near-perfect Top-10 and Top-20 accuracy suggests that the embedding space preserves meaningful local semantic structure. This level of performance is sufficient for image retrieval scenarios or clinical decision-support systems, where presenting a small set of highly relevant candidates is often more valuable than producing a single definitive prediction.

C. THE ROLE OF SIMILARITY METRICS IN EMBEDDING EVALUATION

In connection with the choice of similarity metric, the use of cosine similarity in this study proves to be well aligned with the nature of the task. When diseases exhibit highly similar morphological characteristics and differ primarily in subtle microstructural patterns, the direction of the embedding vector carries more meaningful information than its absolute magnitude.

Cosine similarity directly evaluates the relative relationship between images within the embedding space, thereby emphasizing the model's ability to cluster cases according to underlying pathology. By focusing on angular separation rather than vector length, it highlights whether semantically similar lesions are oriented in comparable directions in high-dimensional space—an assumption consistent with contrastive representation learning.

The experimental results further demonstrate that, under a fixed similarity metric, performance differences across models reflect genuine variations in embedding space quality rather than artifacts introduced by metric selection. In other words, because all models are evaluated using the same similarity function, the observed performance gaps can be attributed to differences in representation learning capability rather than computational bias.

This finding strengthens the internal validity of the evaluation framework and enables a fair comparison between different backbone architectures. It confirms that the reported performance trends are driven by model design and domain alignment, rather than by inconsistencies in similarity measurement.

D. IMPLICATIONS FOR FINE-TUNING PIPELINE DESIGN

From an application-oriented perspective, the findings indicate that selecting a backbone architecture for a dermatological fine-tuning pipeline should not rely solely on immediate classification accuracy. Instead, greater emphasis should be placed on the quality of the initial embedding space. Dermatology-specific models, with semantically structured and well-organized embedding spaces, provide a more robust foundation for subsequent fine-tuning stages—particularly in scenarios where data are limited or class imbalance is present.

Moreover, evaluating models through an image retrieval framework enables the detection of subtle representational differences that traditional classification settings may fail to capture. Retrieval-based evaluation probes the internal geometry of the embedding space rather than only the final decision boundary, making it more sensitive to fine-grained semantic structure.

These observations suggest that retrieval tasks and metric learning approaches should be considered essential intermediate steps in the development of dermatological AI systems. Rather than focusing exclusively on final classification outputs, integrating retrieval-based evaluation and similarity-aware training can lead to more interpretable, stable, and clinically meaningful representations.

V. CONCLUSIONS

This study focuses on evaluating the quality of embedding spaces produced by different models when applied to dermatological images within the infectious disease category, using a cosine similarity–based image retrieval framework. Unlike approaches that involve model training or fine-tuning, this study employs the models in their original, frozen state as fixed feature extractors. This design aims to faithfully reflect the intrinsic representational capacity of each architecture in a dermatological setting characterized by high morphological overlap.

The experimental results demonstrate that embedding quality is strongly influenced by the training data domain of each model. In terms of pure feature extraction performance, ultra-large general-purpose models achieve the highest absolute scores. However, dermatology-specific models—particularly DermLIP—exhibit superior performance-to-computation efficiency in organizing the embedding space according to pathological structure, even when all evaluated diseases belong to the infectious group and share relatively similar clinical manifestations. While large-scale general-purpose models may reach competitive performance on certain Top-k metrics, they still show limitations in separating subtle micro-morphological distinctions specific to dermatology. Meanwhile, general medical models, despite being trained on biomedical data, do not demonstrate a clear advantage in this context, likely due to their broad training scope and lack of specific focus on skin lesion imagery.

Restricting the dataset exclusively to infectious diseases—conditions that often occur in similar anatomical regions and share overlapping visual characteristics in terms of color, texture, and lesion morphology—substantially increases task difficulty. In this setting, high Top-k performance at larger k values reflects the preservation of local semantic structure within the embedding space. Such behavior is particularly well suited to image retrieval and clinical decision-support scenarios, rather than rigid single-label classification.

Overall, the study shows that evaluating embedding quality through an image retrieval paradigm is an effective and practical approach for comparing vision–language backbone architectures in dermatology. The findings provide important empirical evidence for backbone selection in subsequent fine-tuning pipelines and highlight the critical role of domain

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

alignment over mere model scale or raw classification accuracy.

REFERENCES

[1] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[2] Q. Chen et al., "A survey of medical vision-and-language applications and their techniques," arXiv preprint arXiv:2411.12195, 2024.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Machine Learning (ICML), vol. 119, pp. 1597–1607, 2020.

[4] R. Daneshjou, K. Vodrahalli, and R. A. Novoa, "Disparities in dermatology AI performance on a diverse, curated clinical image set," arXiv preprint arXiv:2203.08807, 2022.

[5] DermNet NZ, "Dermatology image dataset," [Online]. Available: https://dermnetnz.org/dermatology-image-dataset

[6] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[7] M. Groh, C. Harris, and L. Soenksen, "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1820–1828, 2021.

[8] F. Khun Jush et al., "Medical image retrieval using pretrained embeddings," arXiv preprint arXiv:2311.13547, 2023.

[9] H. H. Lee, A. Santamaria-Pang, and J. Merkow, "Region-based contrastive pretraining for medical image retrieval with anatomic query," arXiv preprint arXiv:2305.05598, 2023.

[10] [10] A. Mehta, R. Guo, and M. Engel, "Deep learning image processing models in dermatopathology," Diagnostics, vol. 15, no. 19, p. 2517, 2025.

[11] L. Nguyen, "SD-198 skin disease dataset," [Online]. Available: https://www.kaggle.com/datasets/longngzzz/sd-198

[12] OpenAI, "CLIP ViT-Base-Patch32," HuggingFace Model Card, 2021.

[13] OpenAI, "CLIP ViT-Large-Patch14," HuggingFace Model Card, 2021.

[14] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in Proc. ACM Symp. Applied Computing, pp. 1232–1237, 2004.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021.

[16] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," arXiv preprint arXiv:1902.07208, 2019.

[17] M. Rashad, I. Afifi, and M. Abdelfatah, "RbQE: An efficient method for content-based medical image retrieval based on query expansion," Journal of Digital Imaging, vol. 36, pp. 1248–1261, 2023.

[18] A. Sellergren et al., "MedGemma technical report (MedSigLIP)," arXiv preprint arXiv:2507.05201, 2025.

[19] K. Shahhosseini, "RCLIP: CLIP model fine-tuned on radiology images and their captions," HuggingFace Model Card, 2023.

[20] P. Shrestha et al., "Medical vision language pretraining: A survey," arXiv preprint arXiv:2312.06224, 2023.

[21] vinid, "PLIP: Pathology language-image pretraining," HuggingFace Model Card, 2023.

[22] D. Wen et al., "From data to diagnosis: Skin cancer image datasets for artificial intelligence," Clinical and Experimental Dermatology, vol. 49, no. 7, pp. 675–685, 2024.

[23] H. Xu, S. Xie, and X. E. Tan, "Demystifying CLIP data (MetaCLIP)," arXiv preprint arXiv:2309.16671, 2023.

[24] S. Yan, M. Hu, and Y. Jiang, "Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2025.

[25] S. Yan, X. Li, and D. Mo, "A vision-language foundation model for zero-shot clinical collaboration and automated concept discovery in dermatology," arXiv preprint arXiv:2602.10624, 2026.

[26] S. Yan, Z. Yu, and C. Primiero, "A multimodal vision foundation model for clinical dermatology," Nature Medicine, 2025.

[27] S. Yan, Z. Yu, C. Primiero, and C. Vico-Alonso, "DermLIP: A multimodal vision foundation model for clinical dermatology," Nature Medicine, pp. 1–12, 2025.

[28] Y. Yang, M. Gandhi, and Y. Wang, "A textbook remedy for domain shifts: Knowledge priors for medical image analysis," arXiv preprint arXiv:2405.14839, 2024.

[29] K. You, "Semantics at an angle: When cosine similarity works until it doesn't," arXiv preprint arXiv:2504.16318, 2025.

[30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training (SigLIP)," arXiv preprint arXiv:2303.15343, 2023.