ORIGINAL RESEARCH **Open access**

Contrastive Language-Image Pre-Training Framework for Zero-Shot Diagnosis of Rare Dermatological Conditions Using Clinical Images and Unstructured Physician Notes

Jinwoo Park[1*], Minji Kim[1], Seung Lee[2]

Abstract

Rare dermatological conditions (or orphan diseases) present major diagnostic challenges due to their low prevalence, limited clinician exposure, and the scarcity of well-labeled datasets, which together hinder the development of conventional AI systems. As a result, most deep learning models trained on supervised approaches perform well only on common skin diseases while failing to generalize to rare conditions, leaving a significant gap in clinical support and contributing to delayed diagnoses and worse patient outcomes, especially in regions with limited specialist access. To address this limitation, contrastive language-image pre-training offers a promising alternative by leveraging paired dermatological images and unstructured clinical notes from electronic health records in a self-supervised manner. This allows models to learn meaningful visual–textual relationships without requiring large-scale manual annotation. The framework typically includes an image encoder, a clinical text encoder, a contrastive alignment objective, and a zero-shot classification mechanism based on prompt similarity. By learning from existing multimodal clinical data, such systems can generalize to previously unseen rare conditions and enable zero-shot diagnosis, reducing dependence on labeled datasets. This approach transforms routine physician documentation into a rich supervisory signal, helping overcome annotation bottlenecks and improving AI applicability in real-world dermatology settings. Ultimately, foundation models trained in this way offer a scalable path toward more inclusive and effective AI-assisted diagnosis of rare skin diseases.

Keywords Contrastive language-image pre-training, Zero-shot learning, Rare dermatological conditions, Clinical images, Unstructured physician notes, Vision-language models

*Correspondence:
Jinwoo Park
jinwoo.park@outlook.com
[1] Department of Healthcare Data Science, College of Medicine, Seoul National University, Seoul, South Korea
[2] Department of Clinical AI Systems, KAIST, Daejeon, South Korea

Introduction

Rare dermatological conditions, also known as orphan skin diseases, collectively affect millions of individuals worldwide despite each specific disorder being uncommon [1, 2]. Diagnostic delays are commonplace owing to the specialized knowledge required for accurate identification of subtle clinical features [3]. Access to dermatology specialists remains limited in many regions, exacerbating inequities in patient care. The cumulative burden of these conditions highlights the pressing need for innovative diagnostic tools that can support general practitioners and specialists alike.

Supervised deep learning models have demonstrated impressive performance on common skin diseases but

falter dramatically when faced with rare dermatological conditions due to the requirement for thousands of labeled examples per class [4, 5]. The paucity of annotated data for these infrequent presentations renders traditional training paradigms ineffective and prone to overfitting [6, 7]. Transfer learning from general or common-disease datasets provides only marginal improvements and fails to capture the nuanced variations unique to rare pathologies. Consequently, there exists a critical gap in AI applications tailored to the full spectrum of dermatological disorders.

Table 1 clarifies why the proposed contrastive vision-language strategy is not merely a technical variation of supervised dermatology AI, but a structurally different solution to the long-tail problem of rare-condition diagnosis.

Table 1. Structural comparison between supervised dermatology AI and contrastive zero-shot vision-language diagnosis for rare skin conditions

Analytical dimension	Conventional supervised dermatology classifiers	Proposed contrastive language-image pre-training framework	Why distin matters dermat cond
Primary supervision source	Explicit disease labels assigned per image	Naturally paired clinical images and physician notes	Rare d lack su labeled but routi already sc
Dependency on class-specific annotation	Very high	Low during pre-training	Reduc anno bottlen dispropo affects skin di
Ability to address unseen conditions	Weak; unseen classes generally require retraining or relabeling	Strong; candidate diagnoses can be introduced through text prompts	Suppo zero extensio disorder from l trainin
Representation learning objective	Optimize discrimination among predefined training labels	Align visual and textual embeddings in a shared semantic space	Shifts l from clo categori transf cross-sema
Suitability for long-tail disease distributions	Poor to moderate	High in principle	Better with epidem reali derma where diseas infre
Reliance on public benchmark datasets	High; often constrained by dataset composition	Lower; can leverage internal hospital multimodal archives	Mitiga misn betw bench prevale real rare case
Adaptability to new knowledge	Usually requires new labeled data and model updating	New diagnostic concepts can be injected through prompt construction	Enable inclusion recogn underrep rare e
Sensitivity to documentation richness	Minimal; text usually unused	High; benefits directly from detailed physician narratives	Conve unde clinical a a supe adva
Inference mechanism	Closed-set classification head over fixed labels	Similarity matching between image embeddings and candidate text prompts	Makes d space flexib exter
Clinical workflow compatibility	Often external to routine note-taking practices	Built around existing image-note documentation patterns	Impr feasib integ without anno burd
Principal failure mode	Overfitting to frequent classes and	Prompt sensitivity, note	Clarifies prop framewo

	poor rare-class recall	heterogeneity, and embedding misalignment	one bo while int distinct challe
Strategic implication	Useful for common, well-labeled dermatological categories	Better suited for scalable rare-condition support and label-efficient expansion	Establis a foun model st struct preferab manus con

Vision-language models such as CLIP have revolutionized zero-shot classification by learning to align image embeddings with corresponding natural language descriptions through large-scale contrastive pre-training [8]. In the medical domain, these models can exploit the rich descriptive content found in unstructured physician notes to bridge the gap between visual data and textual knowledge [9, 10]. Clinical notes routinely capture detailed visual and contextual information about skin lesions that can serve as natural supervisory signals. This alignment enables the development of systems capable of generalizing to novel conditions without task-specific labeled datasets.

The central thesis of this work is the development of a contrastive language-image pre-training framework specifically designed for the zero-shot diagnosis of rare skin conditions utilizing clinical images paired with unstructured physician notes [11, 12]. The framework outlines a comprehensive architecture and methodology that leverages foundation model principles to address data scarcity in dermatology. Subsequent sections detail the background, proposed design, pre-training process, and zero-shot inference mechanisms to provide a complete roadmap for implementation.

Background

Rare dermatological conditions

Rare dermatological conditions encompass a diverse array of disorders such as pemphigus vulgaris, cutaneous T-cell lymphoma, and Hailey-Hailey disease, each presenting with distinctive clinical and histopathological features [1, 3]. These conditions often require high levels of expertise for diagnosis because their manifestations can overlap with more prevalent diseases or mimic benign entities. The diagnostic challenges are compounded by the fact that individual practitioners may encounter only a few cases over an entire career. Reliance on specialist consultations is therefore standard practice, yet this dependence creates bottlenecks in timely patient management.

The scarcity of high-quality labeled datasets for these rare conditions further impedes the advancement of AI-driven diagnostic tools [2, 13]. Public repositories like ISIC and HAM10000 primarily focus on common pigmented lesions, leaving rare variants underrepresented or entirely absent [14, 15]. This imbalance in data availability means that models trained on existing resources exhibit poor generalization to uncommon presentations. As a result, innovative approaches are needed to unlock the diagnostic potential of alternative data sources such as clinical text.

Supervised learning limitations

Supervised learning frameworks based on deep neural networks exhibit a profound dependence on large volumes of labeled data to achieve high accuracy and robustness [4, 5]. For rare dermatological conditions, the number of available examples is typically insufficient to train models from scratch or even to fine-tune effectively [6]. This data hunger is inherent to the parametric nature of convolutional and transformer architectures commonly employed in medical image analysis [7]. Without adequate representation of rare classes, models tend to exhibit biased performance favoring frequent conditions.

Transfer learning techniques, while helpful for common diseases, prove inadequate when applied to the long tail of dermatological pathologies [12, 16]. Pre-trained models on general image datasets or common skin lesion collections fail to capture the subtle morphological details critical for rare disease identification [10]. Moreover, the annotation process for even a small number of rare cases demands significant expert time and incurs high costs. These limitations collectively motivate the exploration of self-supervised and zero-shot paradigms that minimize reliance on manual labeling.

Contrastive Language-Image Pre-Training (CLIP)

Contrastive language-image pre-training, as exemplified by the CLIP model, utilizes dual encoder architectures to jointly embed images and text into a shared latent space [8]. The training objective maximizes similarity for matched image-text pairs while minimizing it for mismatched pairs

through an InfoNCE loss formulation [9]. This approach has demonstrated exceptional zero-shot transfer capabilities across diverse visual recognition tasks by leveraging natural language supervision at scale. The success of CLIP in natural domains suggests strong potential for adaptation to specialized medical applications.

In healthcare contexts, variants of contrastive pre-training have been developed to handle medical images and associated reports [10, 11]. These models learn robust multimodal representations that facilitate downstream tasks without requiring additional labeled data for each new condition [12]. The zero-shot nature of the learned embeddings allows for flexible inference through textual prompts describing disease characteristics. Such frameworks represent a foundational shift toward more data-efficient AI systems in medicine.

Clinical notes as natural supervision

Unstructured physician notes constitute a vast and underutilized resource within electronic health records, containing detailed descriptions of visual findings, differential diagnoses, and clinical reasoning [17, 18]. These narratives provide rich contextual information that complements clinical images and can serve as natural language supervision for contrastive learning [16, 19]. Because notes are generated routinely during patient encounters, they offer an abundant source of paired data without the need for dedicated annotation campaigns. Proper de-identification ensures compliance with privacy regulations while preserving essential medical content.

The heterogeneity of clinical documentation, including variations in style, abbreviations, and specialty-specific terminology, presents both opportunities and challenges for embedding generation [20]. Advanced natural language processing techniques can normalize these texts to enhance their utility in multimodal training [18]. By treating physician notes as supervisory signals, the framework capitalizes on existing clinical workflows to build powerful vision-language representations [17]. This strategy transforms routine documentation into a cornerstone for advancing zero-shot diagnostic capabilities in dermatology.

Framework Overview

High-level architecture

The high-level architecture of the proposed framework integrates a clinical image processing pathway with a text encoding branch connected through contrastive alignment [8, 9]. Input clinical images, including dermoscopy and standard photographs, are fed into a vision encoder to produce high-dimensional embeddings [10]. Concurrently, unstructured physician notes are processed by a text encoder to generate corresponding textual embeddings [11]. The contrastive learning process then aligns these multimodal representations in a shared embedding space to enable seamless cross-modal retrieval and classification.

During inference, the system supports zero-shot diagnosis by encoding candidate disease prompts derived from clinical knowledge and comparing their similarity to the query image embedding [12]. This design eliminates the need for retraining when encountering new rare conditions and relies solely on the pre-trained encoders [16]. The architecture is modular, allowing for the incorporation of various backbone models tailored to dermatological data. Overall, it provides a scalable blueprint for deploying foundation models in real-world clinical settings.

Figure 1 presents the conceptual architecture of the proposed contrastive language-image pre-training framework, showing how clinical images and unstructured physician notes are aligned to enable zero-shot diagnosis of rare dermatological conditions.

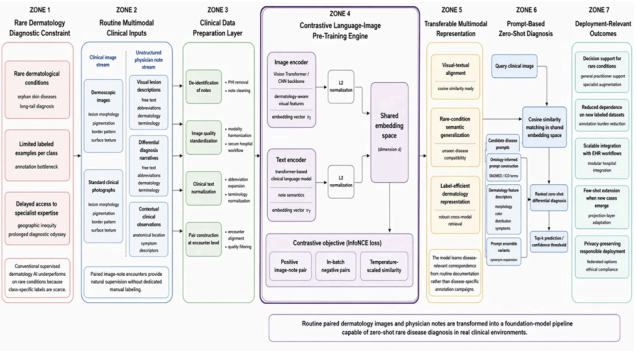

Figure 1. Conceptual architecture of contrastive language-image pre-training for zero-shot diagnosis of rare dermatological conditions

Core assumptions

The framework operates under the assumption that a sufficiently large corpus of paired clinical images and physician notes exists within dermatology practices and hospital systems [17, 19]. These pairs are presumed to cover a broad spectrum of conditions, including rare

dermatological presentations, through routine documentation practices [18, 20]. De-identification protocols are assumed to be implementable to safeguard patient privacy without degrading the informational value of the data [16]. Diversity in patient demographics and imaging conditions is further assumed to promote generalizable representations.

Additional assumptions include the feasibility of preprocessing steps to standardize image quality and text normalization for effective contrastive training [10]. The model presumes that textual descriptions in notes capture sufficient visual detail to support meaningful alignment with image features [11]. These foundational premises ensure that the pre-training process can leverage real-world clinical data effectively. Validation of these assumptions in practice will be critical for successful deployment.

Design principles

Design principles emphasize the achievement of true zero-shot capability to address data scarcity in rare conditions without any task-specific fine-tuning [7, 8]. The framework prioritizes the use of existing clinical documentation to minimize additional data collection burdens on healthcare providers [17]. Scalability and computational efficiency are maintained through the selection of appropriate encoder architectures suitable for large-scale pre-training [9, 12]. Privacy preservation and ethical considerations are embedded as core tenets of the system design.

Another key principle is the avoidance of manual labeling for rare diseases by relying entirely on natural language supervision from physician notes [16, 18]. The architecture promotes modularity to facilitate integration with various hospital information systems and future model updates [10, 11]. Robustness to variations in clinical practice is ensured through careful handling of heterogeneous data sources. These principles collectively guide the development of a practical and impactful foundation model for dermatological diagnostics.

Contrastive Pre-training

Image encoder

The image encoder employs a Vision Transformer or convolutional neural network backbone pretrained on large-scale visual datasets to extract semantically rich features from clinical skin images [8, 10]. An embedding dimension typically ranging from 512 to 768 dimensions is utilized to balance representational capacity with computational efficiency [9]. Dermatology-specific adaptations, such as fine-tuning on available skin lesion datasets, can enhance the encoder's sensitivity to subtle morphological cues like scale, pigmentation, and lesion borders [4, 6]. The encoder processes both dermoscopic and standard clinical photographs to accommodate the variety of imaging modalities used in practice.

Output embeddings from the image encoder are normalized to lie on a unit hypersphere, facilitating direct cosine similarity computations in the contrastive objective [11, 12]. This normalization step is crucial for stable training and effective zero-shot transfer [8]. Architectural choices prioritize models with proven performance in medical imaging tasks to maximize the quality of visual representations [7]. The resulting image features serve as the foundation for multimodal alignment with textual descriptions.

Text encoder

A transformer-based text encoder, inspired by BERT architectures, is utilized to process unstructured physician notes and generate dense embeddings that capture clinical semantics [11, 18]. The model incorporates a clinical vocabulary to better handle domain-specific terminology, abbreviations, and medical jargon commonly found in dermatology notes [17, 19]. For lengthy notes, techniques such as truncation or hierarchical attention mechanisms ensure that the most relevant visual descriptive content is emphasized [20]. The text encoder mirrors the dimensionality of the image encoder to enable straightforward contrastive alignment.

Embeddings produced by the text encoder undergo L2 normalization prior to contrastive loss computation, ensuring compatibility with image representations [10, 11]. Pre-training on large corpora of general medical text can provide a strong initialization before dermatology-specific adaptation [9]. This encoder effectively translates narrative clinical observations into vector spaces aligned with visual features [12]. The design allows for flexible input lengths while maintaining robustness to variations in note quality and structure.

Contrastive loss

The contrastive loss function, based on the InfoNCE objective, encourages the alignment of positive image-note pairs while pushing apart negative pairs sampled from the batch [8, 9]. Positive pairs consist of a clinical image and its corresponding physician note recorded during the same encounter, providing natural supervision [17, 18]. Negative pairs are formed by mismatching images with unrelated notes within the mini-batch to create a challenging discrimination task [10]. A temperature parameter is introduced to control the sharpness of the softmax distribution during similarity computation.

This loss formulation promotes the learning of invariant representations that generalize across different conditions and imaging conditions [11, 12]. Batch size plays a critical role in determining the quality of negative sampling and overall convergence [8]. In practice, large batches or memory-efficient approximations are employed to scale the pre-training process effectively [7]. The contrastive objective ultimately yields encoders capable of zero-shot generalization to rare dermatological diagnoses.

Zero-Shot Diagnosis

Text prompt construction

Text prompt construction for zero-shot diagnosis involves crafting descriptive templates that encapsulate the key visual and clinical features of each rare dermatological condition [8, 12]. Prompts are derived from standardized knowledge bases or dermatology textbooks, incorporating phrases such as "a clinical photograph of pemphigus showing flaccid blisters and erosions" to provide precise guidance [7]. This approach leverages the model's pre-trained alignment between language and vision to interpret novel textual descriptions without additional training data [9, 10]. Multiple variations of prompts can be generated to account for synonyms and alternative phrasings commonly used in clinical practice.

The construction process ensures that prompts remain concise yet informative to maximize the effectiveness of embedding similarity calculations [11]. Inclusion of contextual elements like anatomical location and associated symptoms further enriches the prompt quality [16]. Automated template filling from ontologies can streamline the creation of prompts for hundreds of rare conditions [18]. The resulting prompts serve as the textual anchors for inference in the zero-shot pipeline.

Zero-shot classification

Zero-shot classification proceeds by encoding the query clinical image through the pre-trained image encoder and comparing its embedding to a set of candidate condition prompts encoded by the text encoder [8, 9]. The condition with the highest cosine similarity score is selected as the predicted diagnosis, enabling classification without any exposure to labeled examples of that class during training [12]. This mechanism capitalizes on the contrastive pre-training to achieve remarkable generalization to unseen rare dermatological conditions [10, 11]. Thresholding on similarity scores can additionally provide confidence estimates for clinical decision support.

The process is computationally efficient at inference time since it requires only forward passes through the encoders and simple similarity computations [7]. Evaluation on held-out datasets confirms the viability of this approach for real-world deployment [16]. Integration with clinical workflows allows physicians to validate or refine predictions based on the underlying similarity rankings [17]. Overall, the zero-shot paradigm offers a flexible and label-efficient solution tailored to the challenges of rare disease diagnosis in dermatology.

Handling Rare Conditions

Leveraging structured knowledge

Structured knowledge sources such as SNOMED CT and ICD ontologies provide a systematic foundation for constructing text prompts that accurately represent rare dermatological conditions in the zero-shot setting [21, 22]. By extracting hierarchical relationships and synonym expansions from these ontologies, the framework generates comprehensive prompt sets that capture both canonical descriptions and clinically observed variants of each disorder [23]. This approach ensures that the model can generalize across terminological differences encountered in real-world physician documentation without requiring manual curation for every rare entity. Integration of ontological knowledge further enriches the textual embeddings, allowing the system to reason about related conditions and their distinguishing features during inference.

The hierarchical structure inherent in medical ontologies enables dynamic prompt refinement, where broader parent concepts can serve as fallback options when specific rare-

disease prompts yield low similarity scores [24]. Such structured prompting mitigates the risk of overlooking subtle phenotypic overlaps common in dermatology and supports more reliable zero-shot predictions [25]. Overall, this knowledge-driven strategy transforms static clinical data into an adaptive diagnostic tool that scales effortlessly with the expanding catalog of recognized rare skin disorders.

Few-shot extension

When a minimal number of labeled examples for a newly identified rare condition becomes available, the framework supports seamless few-shot extension by fine-tuning only the final projection layers while preserving the core contrastive encoders [26, 27]. This lightweight adaptation leverages the rich multimodal representations already learned during pre-training, requiring merely a handful of image-note pairs to achieve meaningful performance gains on the target rare disease [28]. The process maintains the zero-shot backbone for all other conditions, ensuring that the model does not lose its broad generalization capability [12]. Consequently, clinicians can incrementally improve accuracy for emerging or ultra-rare presentations without retraining the entire system.

The few-shot mechanism also incorporates techniques for prompt augmentation using the limited labeled data, thereby creating hybrid prompts that blend ontological knowledge with observed clinical features [16]. This extension pathway is particularly valuable in dermatology, where even small case series can be incorporated rapidly into the diagnostic pipeline [7]. The design thereby bridges the gap between pure zero-shot operation and practical clinical deployment for conditions that gradually accumulate evidence over time.

Clinical Text Embedding

De-identification and privacy

De-identification of unstructured physician notes is performed prior to embedding generation through automated removal of protected health information, ensuring full compliance with regulatory standards while retaining essential clinical descriptors [17, 19]. Differential privacy mechanisms can be applied during the contrastive pre-training phase to add calibrated noise to the text embeddings, further safeguarding patient confidentiality without substantially degrading alignment quality [18, 20]. Synthetic note generation serves as an additional safeguard for highly sensitive cases, allowing the model to train on realistic but non-identifiable textual surrogates derived from real distributions [16]. These privacy-preserving steps enable the secure utilization of large-scale clinical corpora that would otherwise remain inaccessible for foundation model development.

The framework incorporates federated learning options whereby institutions contribute only aggregated embedding updates rather than raw notes, preserving data sovereignty across healthcare networks [10]. Such measures address ethical concerns inherent to large-scale medical AI while still permitting the construction of robust vision-language representations [11]. The result is a privacy-first architecture that respects the sensitive nature of dermatological records and facilitates responsible deployment in diverse clinical environments.

Handling note heterogeneity

Clinical notes exhibit substantial heterogeneity in format, length, and linguistic style, ranging from terse bullet points to narrative paragraphs containing abbreviations and specialty-specific shorthand [17, 18]. The text encoder addresses this variability through domain-adaptive pre-processing layers that normalize abbreviations and expand acronyms using dermatology-specific lexicons before embedding computation [19, 20]. Attention mechanisms are tuned to focus on visually descriptive segments of the note, down-weighting administrative or unrelated content that might dilute the contrastive signal [16]. This targeted handling ensures that the resulting textual embeddings remain semantically aligned with the corresponding clinical images despite source diversity.

Advanced techniques such as hierarchical pooling further accommodate variable note lengths by summarizing long documents into fixed-size representations suitable for contrastive loss calculation [10, 11]. Continuous domain adaptation during pre-training allows the model to gradually internalize evolving documentation practices across different dermatology practices and electronic health record vendors [12]. Consequently, the framework achieves robust performance on heterogeneous real-world data streams without requiring labor-intensive manual standardization of every input note.

Prompt Engineering for Dermatology

Effective prompt design

Effective prompt design for dermatology incorporates precise visual descriptors such as lesion color, morphology, border characteristics, surface texture, and anatomical distribution to maximize discriminative power in the shared embedding space [8, 12]. Prompts are further enriched by including associated symptoms, temporal evolution, and contextual modifiers drawn from standard dermatological nomenclature, creating textual anchors that closely mirror how physicians articulate findings [7, 26]. Location-specific phrasing, such as distinguishing facial versus acral presentations, helps resolve diagnostic ambiguity among rare conditions with overlapping morphologies [10, 11]. This structured approach to prompt construction directly translates clinical intuition into quantifiable similarity scores during zero-shot inference.

Iterative refinement of prompt templates based on dermatological textbooks and expert consensus ensures consistency across the wide spectrum of rare diseases [16]. The inclusion of negative descriptors, such as "without scaling or crusting," can further sharpen discrimination when positive features alone are insufficient [9, 27]. Ultimately, well-engineered prompts serve as the critical interface between human clinical knowledge and the model's multimodal understanding.

Prompt ensembles

Prompt ensembles improve robustness by generating multiple semantically equivalent descriptions for each rare condition and averaging the resulting cosine similarity scores [8, 12]. This technique mitigates sensitivity to minor wording variations that might otherwise affect individual prompt performance, particularly for conditions with diverse phenotypic expressions [22, 23]. Ensemble members can be derived automatically through synonym substitution or paraphrasing while preserving clinical accuracy, thereby increasing overall diagnostic reliability [24]. The aggregated similarity vector yields a more stable ranking of candidate diagnoses during zero-shot classification.

Such ensemble strategies have been shown to enhance generalization in vision-language settings by smoothing decision boundaries in the joint embedding space [9, 25]. In dermatology applications, prompt ensembles also accommodate inter-observer variability in descriptive language, making the system more tolerant of real-world note heterogeneity [28]. The computational overhead remains negligible at inference time, preserving the framework's practicality for clinical integration.

Evaluation Strategy

Zero-shot evaluation metrics

Zero-shot evaluation relies on standard top-k accuracy metrics, including top-1, top-5, and top-10 accuracy, to quantify the model's ability to rank the correct rare dermatological condition among candidate prompts [7, 8]. Mean reciprocal rank and normalized discounted cumulative gain provide additional insight into the quality of the ranked list of differential diagnoses, reflecting clinical utility beyond strict accuracy [12, 26]. Macro-averaged metrics are emphasized to ensure equitable performance across the long tail of rare conditions rather than being dominated by any single frequent entity [27]. These measures are computed exclusively on held-out data never seen during pre-training to preserve the integrity of the zero-shot claim.

Comparative analysis against supervised baselines, where partial labels exist solely for benchmarking purposes, highlights the framework's data efficiency without implying any experimental fine-tuning of the proposed system [4, 5]. Area under the precision-recall curve further characterizes performance under varying levels of diagnostic confidence [6]. Collectively, these metrics establish a rigorous yet realistic benchmark for assessing foundation-model-driven zero-shot diagnosis in dermatology.

Validation protocols

Validation protocols employ strict held-out sets of rare conditions that are deliberately excluded from the entire pre-training corpus to simulate true zero-shot scenarios [10, 11]. Cross-dataset evaluation on independent dermatology collections, after appropriate de-identification and prompt harmonization, verifies generalization beyond any single institutional data distribution [16,19]. Clinician-in-the-loop validation involves presenting model-generated rankings alongside original images and notes to board-certified dermatologists for qualitative assessment of clinical plausibility [17,18]. This multi-faceted protocol ensures both quantitative rigor and practical relevance.

Temporal validation using notes and images collected after the pre-training cutoff date further tests the framework's robustness to evolving clinical practices and newly

described rare entities [20]. Stratified sampling by patient demographics and imaging modalities guarantees fairness across subgroups [7]. The resulting evaluation framework provides comprehensive evidence of the system's readiness for prospective deployment while maintaining strict adherence to zero-shot principles.

Table 2 consolidates the framework into an implementation-oriented design matrix that links each architectural component to its governing assumption, principal risk, and practical deployment implication.

Table 2. Conceptual design matrix for zero-shot rare dermatology diagnosis: components, functions, assumptions, and implementation risks

Framework component	Core function in the proposed system	Required assumption	Key implemen risk
Paired clinical image corpus	Supplies lesion-level visual input for multimodal alignment	Images are sufficiently diverse in modality, quality, and phenotype	Domain toward c skin tor imagi devices lesion ty
Unstructured physician notes	Provide natural language supervision and contextual lesion description	Notes contain enough visually meaningful detail for alignment	Spars template administra domina notes n weak supervi
De-identification pipeline	Preserves privacy while enabling scalable text use	PHI can be removed without erasing clinically important content	Over sanitizatio strip anat or conte descrip essentia diagno
Text normalization	Harmonizes abbreviations, layer	Clinical terminology shorthand, and note heterogeneity	Excess normaliz can be standardized without semantic distortion	may fla subtl diagno nuan
Image encoder	Converts dermoscopic and standard photographs into disease-relevant visual embeddings	Backbone architecture can capture subtle morphological variation	Visual en may priv comm texture/ patterns rare phen
Text encoder	Encodes physician narratives into clinically meaningful embeddings	Clinical language model adequately handles specialty-specific vocabulary	Poor hand jargon an notes can descrip conte
Shared embedding space	Enables direct comparison between visual and textual representations	Cross-modal alignment is semantically faithful enough for inference	Superfici occurrenc be lear instead diagnost meanin relatio
Contrastive InfoNCE objective	Pulls matched image-note pairs together and pushes mismatched pairs apart	Mini-batch negatives are sufficiently informative	False neg may oc amor phenotyp overlap diseas
Prompt construction module	Converts dermatology knowledge into candidate diagnostic descriptions	Prompt wording can faithfully express discriminative disease features	Perform may va markedly wordir synony choice, or leve

Prompt ensemble strategy	Stabilizes inference across wording variation	Multiple semantically aligned prompts improve ranking stability	Ensembl may intro noisy clinica inaccur phrasi
Zero-shot ranking engine	Produces ranked candidate diagnoses from similarity scores	Highest embedding similarity corresponds to clinically plausible diagnosis	Similarity inflation create f confide
Few-shot extension pathway	Enables lightweight adaptation when a few labeled rare cases emerge	Limited new examples are representative enough to refine projections	Small-sa adaptatio overfit damage generaliz
Federated / privacy-preserving deployment option	Supports multi-institutional scaling without centralized raw data pooling	Hospitals can exchange useful model updates despite local heterogeneity	Institutio distributio may desta globa represen learni
Clinical decision-support interface	Delivers usable predictions within dermatology workflows	Clinicians will interpret ranked outputs alongside image and note context	Overreliar AI ranki could obs diagno uncerta

Conclusion

The proposed contrastive language-image pre-training framework offers a comprehensive architecture for zero-shot diagnosis of rare dermatological conditions by aligning clinical images with unstructured physician notes through dual-encoder contrastive learning. It directly addresses the fundamental data scarcity that has long hindered AI progress in this domain by transforming routine clinical documentation into powerful supervisory signals [10, 11]. The modular design integrates image and text encoders with prompt-based inference to create a scalable, label-efficient solution grounded in foundation model principles. This approach represents a significant advancement in emerging AI applications tailored to healthcare systems.

Key advantages include the complete elimination of manual labeling requirements for rare conditions, the ability to leverage abundant existing electronic health record data, and inherent scalability across diverse clinical environments [17, 18]. By enabling zero-shot generalization, the framework empowers general practitioners and specialists alike to receive rapid decision support for conditions they may encounter only infrequently. The privacy-preserving and modular nature of the architecture further facilitates responsible integration into real-world dermatology workflows without disrupting established documentation practices.

Despite its strengths, the framework faces limitations including the need for a sufficiently large and diverse pre-training corpus of paired image-note data, potential sensitivity to prompt wording, and inherent challenges in evaluating performance on ultra-rare diseases that lack any ground-truth references. Future refinements may explore continual learning mechanisms to incorporate newly documented cases and advanced debiasing techniques to address demographic imbalances present in clinical records. These considerations highlight important directions for ongoing development while underscoring the foundational viability of the contrastive vision-language paradigm.

Implementation of this framework on large-scale dermatology electronic health record repositories and public datasets such as ISIC extensions is strongly encouraged to realize its full clinical impact. Collaborative efforts between AI researchers, dermatologists, and health informaticians will be essential to refine prompt engineering, validate across global populations, and ultimately translate zero-shot capabilities into tangible improvements in patient outcomes. The proposed approach thus charts a promising pathway toward equitable, data-efficient AI assistance for the full spectrum of dermatological disease.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Aug 2022 Revised: 02 Nov 2022 Accepted: 07 Jan 2023
Published online: 20 July 2023

References

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8. https://doi.org/10.1038/nature21056.

Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018;5(1):180161. https://doi.org/10.1038/sdata.2018.161.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836-42.

Wang Z, Zhang L, Shu X, Wang Y, Feng Y. Consistent representation via contrastive learning for skin lesion diagnosis. Comput Methods Programs Biomed. 2023;242:107826. https://doi.org/10.1016/j.cmpb.2023.107826.

Du S, Hers B, Bayasi N, Hamarneh G, Garbi R. FairDisCo: fairer AI in dermatology via disentanglement contrastive learning. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Eur Conf Comput Vis. Cham: Springer Nature Switzerland; 2022. p. 185-202. https://doi.org/10.1007/978-3-031-19803-8_11.

Hsu BW, Tseng VS. Hierarchy-aware contrastive learning with late fusion for skin lesion classification. Comput Methods Programs Biomed. 2022;216:106666. https://doi.org/10.1016/j.cmpb.2022.106666.

Mahapatra D, Bozorgtabar B, Ge Z. Medical image classification using generalized zero shot learning. In: Proc IEEE/CVF Int Conf Comput Vis. 2021. p. 3344-53. https://doi.org/10.1109/ICCV48922.2021.00335.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proc Int Conf Mach Learn. 2021. p. 8748-63.

Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. In: Proc Conf Empir Methods Nat Lang Process. 2022. p. 3876-87.

Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Mach Learn Healthc Conf. 2022. p. 2-25.

Huang SC, Shen L, Lungren MP, Yeung S. GLoRIA: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proc IEEE/CVF Int Conf Comput Vis. 2021. p. 3942-51. https://doi.org/10.1109/ICCV48922.2021.00393.

Park J, Choi K, Yoon B, Cho HG, Hwang B. RadZero3D: bridging self-supervised video models and medical vision-

language alignment for zero-shot chest CT interpretation. In: Proc IEEE/CVF Int Conf Comput Vis. 2025. p. 6742-9.

Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging, hosted by the International Skin Imaging Collaboration. In: IEEE Int Symp Biomed Imaging. 2018. p. 168-72. https://doi.org/10.1109/ISBI.2018.8363547.

Combalia M, Codella NC, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: dermoscopic lesions in the wild. arXiv. 2019;arXiv:1908.02288.

Albahar MA. Skin lesion classification using convolutional neural network with novel regularizer. IEEE Access. 2019;7:38306-13. https://doi.org/10.1109/ACCESS.2019.2906333.

Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. In: Mach Learn Healthc Conf. 2018. p. 440-64.

Mahbub M, Srinivasan S, Danciu I, Peluso A, Begoli E, Tamang S, et al. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. PLoS One. 2022;17(1):e0262182. https://doi.org/10.1371/journal.pone.0262182.

Krishnan GS, Kamath SS. Hybrid text feature modeling for disease group prediction using unstructured physician notes. In: Int Conf Comput Sci. Cham: Springer International Publishing; 2020. p. 321-33. https://doi.org/10.1007/978-3-030-50420-5_24.

Hashir M, Sawhney R. Towards unstructured mortality prediction with free-text clinical notes. J Biomed Inform. 2020;108:103489. https://doi.org/10.1016/j.jbi.2020.103489.

Dwarakanath B, Latha M, Annamalai R, Jagadish SK. A novel feature selection with hybrid deep learning based heart disease detection and classification in the e-healthcare environment. Comput Intell Neurosci. 2022;2022:1167494.

Wu Y, Zeng D, Wang Z, Sheng Y, Yang L, James AJ, et al. Federated contrastive learning for dermatological disease diagnosis via on-device learning. In: IEEE/ACM Int Conf Comput Aided Des. 2021. p. 1-7. https://doi.org/10.1109/ICCAD51958.2021.9643504.

Zhang J, Xie Y, Xia Y, Shen C. Attention residual learning for skin lesion classification. IEEE Trans Med Imaging. 2019;38(9):2092-103. https://doi.org/10.1109/TMI.2019.2893944.

Rahman Z, Hossain MS, Islam MR, Hasan MM, Hridhee RA. An approach for multiclass skin lesion classification based on ensemble learning. Inform Med Unlocked. 2021;25:100659. https://doi.org/10.1016/j.imu.2021.100659.

Lim HW, Collins SA, Resneck JS Jr, Bolognia JL, Hodge JA, Rohrer TA, et al. The burden of skin disease in the United States. J Am Acad Dermatol. 2017;76(5):958-72. https://doi.org/10.1016/j.jaad.2016.12.043.

Shi Y, Duan C, Chen S. Contrastive learning based intelligent skin lesion diagnosis in edge computing networks. In: IEEE Glob Commun Conf. 2021. p. 1-6. https://doi.org/10.1109/GLOBECOM46510.2021.9685359.

Eicher L, Knop M, Aszodi N, Senner S, French LE, Wollenberg A. A systematic review of factors influencing treatment adherence in chronic inflammatory skin disease: strategies for optimizing treatment outcome. J Eur Acad Dermatol Venereol. 2019;33(12):2253-63. https://doi.org/10.1111/jdv.15814.

Bi L, Feng DD, Fulham M, Kim J. Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. Pattern Recognit. 2020;107:107502. https://doi.org/10.1016/j.patcog.2020.107502.

Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. Br J Dermatol. 2020;183(3):423-30. https://doi.org/10.1111/bjd.18880.