A Transformer-Based CBIR System for Cutaneous Ulcer Analysis with LLM Assistance

Gabriel B. dos Santos[1], Ana Elisa Jorge[1], Agma J. M. Traina[1], Mirela T. Cazzolato[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400 - Centro, 13566-590 – São Carlos, SP – Brazil

gabrielbarbosadossantos@usp.br, anajorge@icmc.usp.br,
agma@icmc.usp.br, mirela@icmc.usp.br

Abstract. *Chronic dermatological ulcers are a public health concern, with clinical assessment subject to inter-observer variability. Content-Based Image Retrieval (CBIR) systems offer decision support, yet Transformer-based feature extractors remain underexplored in this domain. This work presents a CBIR system augmented with a Large Language Model (LLM) layer that summarizes retrieved cases in natural clinical language. To inform its configuration, we compare five Transformer extractors against handcrafted descriptors and a random reference on a dataset labeled by tissue composition, using Precision@K, MAP, macro-mAP, and NDCG@K with graded Jaccard relevance. The study identifies a configuration with consistent performance across all metrics.*

1. Introduction and related work

Many areas of medicine rely on expert visual assessment for diagnosis, treatment selection, and patient follow-up. Such assessment is subject to inter-observer variability, a condition particularly accentuated among trainees still building the clinical repertoire required to distinguish subtle visual patterns [Forti et al. 2026]. The challenge is especially pronounced in chronic dermatological ulcer care, a growing public health concern: in 2022, chronic wounds affected approximately 10.5 million Medicare beneficiaries in the United States [Sen 2023], and pressure ulcers in particular have shown a consistent rise in global burden over the past decades [GBD 2021 Decubitus Ulcers Collaborators 2025]. Visual complexity compounds the problem: tissue composition is heterogeneous, combining granulation (red, viable), fibrin (yellow, devitalized), necrosis (dark), and callus (thickened borders) in variable proportions [Blanco et al. 2020], hindering both initial categorization and longitudinal follow-up.

Content-Based Image Retrieval (CBIR) systems address this need by returning, for a query image, the most visually similar cases from a database, narrowing the semantic gap between visual content and clinical interpretation. In the pressure ulcer domain, prior work from our research group includes ICARUS [Chino et al. 2018], which uses local descriptors aggregated as Bag-of-Signatures, and the approach of Blanco et al. [Blanco et al. 2020], which combines superpixels with convolutional networks for locally coherent wound representations. The state of the art has since advanced with Transformer-based architectures [Vaswani et al. 2017] and Vision-Language Models (VLMs), which produce semantically rich embeddings. Their application to cutaneous ulcers remains underexplored, and it is unclear whether general-purpose pretraining suffices for the domain or whether biomedical variants offer consistent gains.

Beyond retrieval itself, the professional must translate visual patterns into clinical terms, a step that is particularly demanding for those in training. Large Language Models (LLMs) can serve as a complementary interpretive layer, summarizing, in natural clinical language, the predominant characteristics of the retrieved set. This work proposes a CBIR system for cutaneous ulcers augmented by such a layer, whose configuration is informed by a comparative study of five Transformer-based extractors against handcrafted (traditional) and random baselines.

The main contributions of this work are: (i) a comparative study of Transformer-based extractors CLIP, BiomedCLIP, SigLIP 2, DINOv2, PMC-CLIP against handcrafted and random baselines in the cutaneous ulcer domain; (ii) identifying a configuration with consistent performance across the evaluated metrics, considering extractor, similarity function, and indexing strategy; and (iii) a functional CBIR system with an LLM-based interpretive layer (Gemini 2.5 Flash) intended to support wound-care trainees.

2. Methodology

Dataset. The dataset (referred to here as ULCER-A) is from [Cazzolato et al. 2021], and has 217 chronic dermatological ulcer images from Hospital das Clínicas de Ribeirão Preto, multi-labeled by tissue composition (G=granulation, F=fibrin, N=necrosis, C=callus) and imbalanced.

CBIR pipeline. The CBIR pipeline adopted in this work follows the classical three-stage structure illustrated in Figure 1. Each database image is processed by a feature extractor that produces an embedding vector representing its visual content. The resulting vectors are organized into an indexing structure that supports efficient querying. Given a query image, its embedding is computed and compared against those in the database via a similarity function, returning the most similar items. An LLM-based interpretive layer (detailed below) is added on top, operating only on the labels and similarities of the retrieved set without affecting the retrieval process itself.

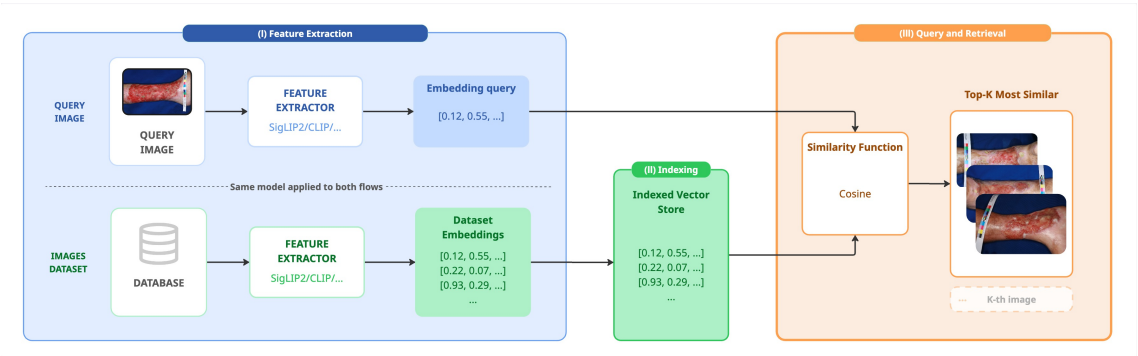

Figure 1. Three-stage CBIR pipeline: feature extraction with the same model applied to both the query image and the database images, indexing of the resulting embeddings, and similarity-based retrieval of the top-K most similar cases.

Feature extractors. We selected five Transformer-based extractors covering three representational paradigms. CLIP [Radford et al. 2021] (ViT-L/14) and SigLIP 2 [Tschannen et al. 2025] (large variant) represent contrastive image-text alignment on general-purpose corpora, with SigLIP 2 incorporating sigmoid loss and self-

distillation techniques that position it as state of the art in zero-shot retrieval. DINOv2 [Oquab et al. 2024] (large) adopts a purely visual self-supervised paradigm, trained without textual alignment on a curated set of approximately 142 million images. BiomedCLIP [Zhang et al. 2023] (ViT-B/16) and PMC-CLIP [Lin et al. 2023](community-released ViT-L/14 checkpoint) are variants of the CLIP architecture pretrained on biomedical corpora extracted from PubMed Central, with 15M and 1.6M image-text pairs, respectively. This composition allows us to analyze both the effect of the pretraining paradigm and the potential gain from biomedical specialization for the target domain.

As low-level baselines, we considered seven traditional (handcrafted) descriptors, each computing a feature vector over the same images as the Transformer extractors: LBP, edge histogram, color layout, scalable color, color structure, texture spectrum, and a normalized color histogram. These descriptors capture isolated aspects of texture, color, and contour, serving as a reference point to measure the effective gain provided by Transformer-based extractors.

Images were processed using each model's official preprocessing routines. Embeddings were used off-the-shelf, without fine-tuning, given the study's focus on comparing the quality of the pretrained representations themselves.

Indexing and similarity-based search. Embedding similarity is computed via cosine distance, with embeddings L2-normalized so that cosine similarity is equivalent to the inner product. Indexing is performed with FAISS [Johnson et al. 2019] using `IndexFlatIP` for exact search, kept for pipeline modularity. Queries return the top-K items with the highest cosine similarity to the query embedding.

Evaluation metrics. Retrieval evaluation under multi-label scenarios, as in the dataset used here, requires a careful definition of relevance. We adopt the Jaccard index between the class sets of the query and the retrieved item: in binary form (strict Jaccard, threshold 0.5) for P@K, R@K, AP, and MAP, avoiding the metric inflation observed under looser criteria in the presence of a dominant class, and as a continuous value for the graded gain of NDCG@K, partially rewarding partial class overlaps.

To mitigate the strong imbalance (class G is present in roughly 74% of the images), we also report macro-mAP per base class: for each class (G, F, N, C), we compute the average AP over the queries whose label contains that class, and take the unweighted mean of the four values. As an additional reference, we compute a random baseline in which, for each query, the ranking of gallery items is drawn as a uniformly random permutation rather than ordered by similarity; all retrieval metrics are then computed over this random ranking. This is repeated via Monte Carlo (100 independent draws per query), and we report the lift of each model as the ratio of its mean metric to the baseline's.

LLM-assisted interpretive layer. The interpretive layer transforms the retrieved set into a descriptive reading expressed in natural clinical language. It is implemented with Gemini 2.5 Flash, accessed via its official API, at a temperature of 0.3, balancing stylistic stability with output fluency.

The prompt is structured in three parts. First, it provides context on the dataset's labeling criterion – tissue composition into granulation, fibrin, necrosis, and callus – and the clinical meaning of each class, anchoring the model's vocabulary to the wound-care domain. Second, it presents the queried case, indicating only its label. Third, it provides

a numbered list of the retrieved neighbors in descending order of similarity, with their respective labels and similarity values. A central design choice is that the model receives only labels and similarity values; no images are sent to the LLM. This grounds the textual analysis on what the retrieval system actually returned, rather than on independent visual reasoning about the lesion, and as a useful side effect allows the interpretive output to be displayed in clinical contexts where exposing patient images would be undesirable.

Response guidelines instruct the model to describe the predominant tissue characteristics of the retrieved set, comment on its internal coherence, and remark on any divergence relative to the queried case's label. As a safeguard, the prompt actively restricts the model from issuing diagnostic statements or recommending treatment, keeping the output as a descriptive reading of the retrieved set rather than an assessment of the queried lesion.

3. Experiments and results

Experimental setup. Experiments were run in Python using PyTorch, `transformers`, and FAISS. We adopt 5-fold cross-validation grouped by `patient_id` and stratified by tissue composition: in each iteration, four folds form the search gallery and the remaining fold provides the queries, ensuring that images from the same patient never appear simultaneously on both sides and preventing information leakage. The random seed is fixed (`seed=42`) and all models share the same folds; each metric is reported as the mean over five folds; standard deviations were below 0.10 for all models and metrics.

Comparative evaluation. Table 1 consolidates the results of the comparative evaluation between Transformer-based extractors, the best handcrafted descriptor, and the random baseline, aggregated over the five folds. Among the Transformers, SigLIP 2 appears at the top of the ranking in every reported metric (macro-mAP of 0.747 and P@1 of 0.798), being the only model to display this pattern. PMC-CLIP, CLIP, and BiomedCLIP occupy close positions to each other, with the relative ordering varying by metric, and DINOv2 falls into the last position among the Transformers, still clearly above the baselines.

Table 1. Comparative evaluation of extractors in the ULCER-A dataset. Each value is the average over five folds. The best results are in bold.

Model	macro-mAP	mAP	P@1	NDCG@1	P@5	NDCG@10
SigLIP 2	**0.747**	**0.821**	**0.798**	**0.618**	**0.773**	**0.549**
PMC-CLIP	0.707	0.804	0.782	0.583	0.737	0.531
CLIP ViT-L/14	0.712	0.797	0.757	0.551	0.740	0.521
BiomedCLIP	0.670	0.792	0.760	0.561	0.742	0.538
DINOv2	0.700	0.782	0.739	0.560	0.721	0.519
Color Structure (Best handcrafted)	0.623	0.734	0.672	0.447	0.660	0.445
Random Baseline	0.578	0.705	0.638	0.432	0.630	0.433

Handcrafted descriptors stay close to the random baseline in macro-mAP (the best of them, *color structure*, reaches 0.623 against 0.578 for the random baseline), indicating that isolated aspects of color, texture, and contour do not capture sufficiently discriminative features for the task.

The discrepancy between micro and macro metrics also deserves attention. In micro-mAP, the separation between Transformers and handcrafted descriptors is subtler.

In macro-mAP, this gap widens, indicating that Transformers gain consistency when performance is balanced across the four tissue classes. This reinforces the relevance of reporting metrics that control for the imbalance effect in clinical datasets like ULCER-A.

The gap also widens at the top of the ranking: in NDCG@1, SigLIP 2 reaches 0.618 against 0.432 for the random baseline (lift of 1.43×), while in NDCG@10 this ratio drops to 1.27×. The advantage of Transformer-based extractors therefore concentrates in the first positions of the ranking, exactly where the operational impact of the system is greatest.

Patient identity analysis. Since ULCER-A contains multiple images per patient across different healing stages, we ask whether retrieval surfaces lesions with comparable tissue composition or other images of the *same* lesion. For each query and K, the intra-patient *lift* is the ratio between the share of top-K neighbors sharing `patient_id` with the query and the share expected under uniform random sampling. Values near 1 indicate no preference for same-patient images; values above 1 indicate systematic intra-patient retrieval. Unlike the patient-grouped split of the comparative evaluation, this analysis uses a within-fold protocol where the query and gallery come from the same fold (excluding the query itself), so that same-patient images are present and the intra-patient tendency can be measured. Results are averaged across folds.

Table 2. Intra-patient *lift* at $K = 1$, 5, and 10. Values above 1 indicate that the model retrieves images of the same patient more often than expected under uniform random sampling. Edge Histogram is reported as the handcrafted descriptor with the highest intra-patient lift.

Model	*lift*@1	*lift*@5	*lift*@10
SigLIP 2	9.89	5.36	3.54
PMC-CLIP	9.55	5.14	3.39
DINOv2	9.02	4.89	3.19
CLIP ViT-L/14	8.68	4.76	3.29
BiomedCLIP	8.62	4.64	3.17
Edge Histogram (Best handcrafted)	1.83	1.04	1.00

Table 2 shows a consistent pattern among the Transformers: high lifts at $K = 1$ (8.6–9.9), decreasing but still substantial at $K = 10$ (3.2–3.5). Handcrafted baselines stay near 1 at all K, showing no such tendency. The pattern holds on the ULCER-A ROI variant (images cropped to the lesion, removing background and adjacent skin), indicating the behavior is driven by lesion features rather than context. Notably, SigLIP 2 — the most consistent extractor in Table 1 — also shows the highest intra-patient lift; both stem from finer-grained visual discrimination and characterize the retrieval behavior of Transformer-based extractors rather than confound the comparison.

In real deployment, where the gallery is not partitioned by patient, this tendency may cause recurring patients to dominate the top-K, impoverishing case diversity. Diversification by `patient_id` at retrieval time can mitigate this without changing the extractor.

System interface. The proposed configuration is delivered as an interactive `Streamlit` application, illustrated in Figure 2. Given a query image, the professional sees the K most

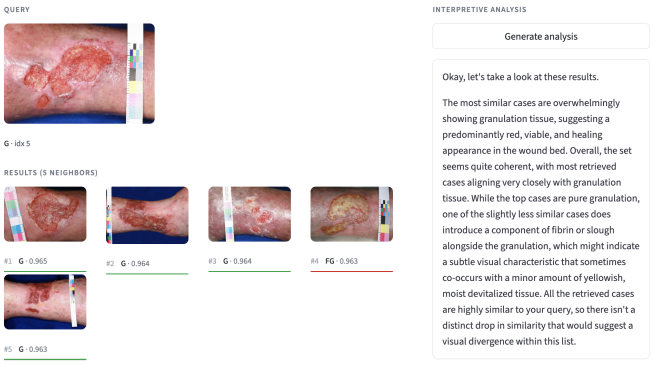

Figure 2. Streamlit interface: the query image and top-K retrieved cases with labels and similarity scores (left), and the interpretive layer's textual synthesis (right).

similar cases (for direct visual validation) together with the textual synthesis generated by the interpretive layer over the retrieved set. Because the LLM operates exclusively on labels and similarity values, the image display and the textual analysis remain independent: in scenarios where direct exposure of referenced patient images is undesirable, the textual output can be presented alone.

4. Conclusion and future work

This work presented a Content-Based Image Retrieval system for cutaneous ulcers, augmented by an LLM-based interpretive layer, with configuration informed by a comparative study of five Transformer-based extractors, handcrafted baselines, and a random baseline on the ULCER-A dataset. The proposed configuration combines SigLIP 2 (large) with cosine similarity, FAISS exact indexing, and Gemini 2.5 Flash as the interpretive layer; SigLIP 2 stood out as the only model at the top of every reported metric, while PMC-CLIP and CLIP performed comparably and can replace it without structural changes. We also identified that Transformer-based extractors tend to retrieve other images of the same lesion at high rates — a behavior absent in handcrafted baselines and relevant to the operational design of the system.

Several limitations condition the results. The dataset comes from a single clinical center and is small in scale, restricting the generalization of the conclusions and precluding the investigation of fine-tuning strategies without substantial overfitting risk. The intra-patient retrieval tendency, while not compromising the relative comparison between models, motivates the inclusion of diversification mechanisms for real clinical use. Future work includes expanding the dataset through multi-center curation, evaluating domain-specific fine-tuning strategies, incorporating patient-level diversification at retrieval time, and validating the system with wound-care trainees through a qualitative study.

Acknowledgements. This work was supported by the São Paulo Research Foundation (FAPESP–grants #2026/11130-2, #2024/15430-5, #2024/13328-9), the Coordination for Higher Education Personnel Improvement (CAPES–Finance Code 001), the National Research Council (CNPq–grants PIBIC #4356/2025), and the University of São Paulo (PRPI 1032 - grant #207).

Origin of the work. This paper reports results obtained from the Scientific Initiation (IC) project of G.B.S.

Use of generative AI. Generative AI tools (LLMs) were used to assist with text structuring and language revision. All scientific content, results, and references are the sole responsibility of the authors.

References

Blanco, G. et al. (2020). A superpixel-driven deep learning approach for the analysis of dermatological wounds. *Computer Methods and Programs in Biomedicine*, 183:105079. DOI: 10.1016/j.cmpb.2019.105079.

Cazzolato, M. T., Ramos, J. S., Rodrigues, L. S., Scabora, L. C., Chino, D. Y. T., Jorge, A. E. S., de Azevedo-Marques, P. M., Traina Jr., C., and Traina, A. J. M. (2021). The UTrack framework for segmenting and measuring dermatological ulcers through telemedicine. *Computer Methods and Programs in Biomedicine*. DOI: 10.1016/j.compbiomed.2021.104489.

Chino, D. Y. T. et al. (2018). ICARUS: Retrieving skin ulcer images through bag-of-signatures. In *Proc. IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 82–87. IEEE. DOI: 10.1109/CBMS.2018.00022.

Forti, J. K., Navarro, T. P., and dos Santos, A. L. (2026). Explainable deep learning for etiological classification of vascular ulcers using a hybrid convolutional neural network-transformer model. *JVS-Vascular Insights*, 4:100419. DOI: 10.1016/j.jvsvi.2026.100419.

GBD 2021 Decubitus Ulcers Collaborators (2025). Global, regional and national burden of decubitus ulcers in 204 countries and territories from 1990 to 2021: a systematic analysis based on the Global Burden of Disease study 2021. *Frontiers in Public Health*, 13:1494229. DOI: 10.3389/fpubh.2025.1494229.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547. DOI: 10.1109/TBDATA.2019.2921572.

Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., and Xie, W. (2023). PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14227 of *Lecture Notes in Computer Science*, pages 525–536. Springer. DOI: 10.1007/978-3-031-43993-3_51.

Oquab, M., Darcet, T., Moutakanni, T., et al. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=a68SUt6zFt.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning (ICML)*. PMLR. https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

Sen, C. K. (2023). Human wound and its burden: Updated 2022 compendium of estimates. *Advances in Wound Care*, 12(12):657–670. DOI: 10.1089/wound.2023.0007.

Tschannen, M. et al. (2025). SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., Poon, H., et al. (2023). BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.