Multimodal Deep Learning for Predicting Aesthetic Outcomes in Breast Cancer Surgery

Mohammad Hossein Zolfagharnasab[1,2]
mohammad.h.zolfagharnasab@inesctec.pt
Nuno Freitas[1,2]
nuno.p.silva@inesctec.pt
Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt
Eduard Bonci[3]
eduard.bonci@research.fchampalimaud.org
Carlos Mavioso[3]
carlos.mavioso@fundacaochampalimaud.pt
Maria J. Cardoso[2,3]
maria.joao.cardoso@fundacaochampalimaud.pt
Hélder P. Oliveira[2,4]
helder.f.oliveira@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Porto, Portugal
[2] Institute for Systems and Computer Engineering, Technology and Science Porto, Portugal
[3] Breast Unit, Champalimaud Foundation Lisboa, Portugal
[4] Department of Computer Science, Faculty of Sciences, University of Porto, Portugal

Abstract

This study introduces a deep learning (DL) multimodal retrieval system to predict post-surgery aesthetic outcomes in breast cancer patients using 2,193 instances combining clinical data and RGB images. We compared four retrieval scenerios, with fine-tuned Vision Transformers (ViT)s achieving up to 73.85% accuracy and 80.62% Adjusted Discounted Cumulative Gain (ADCG). Evaluated on over 20K triplets, our model enhances the prediction of post-surgery aesthetics, helping manage patient expectations and offering broad applications in medical image retrieval.

1 Introduction

As the survival rates of Breast cancer (BrCa) patients have improved to a satisfactory rate in many high-income countries, the focus has shifted towards enhancing post-surgery quality of life (QoL) [1]. Given that current methods for assessing aesthetic outcomes are subjective, with viable semantic gap with clinicians' metrics, there is a clear need for objective retrieval systems using artificial intelligence (AI) to improve patient understanding and aid in case selection by clinicians [2]. To address this, we propose a DL-based multimodal retrieval pipeline to enhance BrCa post-surgery outcome prediction by combining tabular clinical data and images. Additionally, we conduct a comparative study of ViTs and CNNs for feature extraction, a core component of such systems. The code for the implementation of this paper is publicly available on GitHub[1].

2 Mathematical Modelling: Triplet Loss

The triplet loss is a robust technique used in tasks like face recognition, image retrieval, and metric learning to learn embeddings that capture semantic similarity among data points [3]. It involves three entities: a query, a positive sample (similar to the query), and a negative sample (dissimilar to the query). Eq. 1 shows the triplet loss formulation:

$$L(Q,P,N) = \max(0, dist(Q,P) - dist(Q,N) + \alpha), \tag{1}$$

where Q, P, and N are embedding of the query, positive, and negative samples, respectively, $dist(Q,P)$ and $dist(Q,N)$ represent distances (e.g., Euclidean or cosine) between these samples, and α is a margin hyperparameter enforcing a minimum difference between distances for the loss to be zero.

3 Related Work

In recent years, the research on Content-based image retrieval systems (CBIRS) focuses extensively on enhancing feature extraction algorithms [4]. With that in mind, we point the reader to the works of Wang et al. [5], Bhandi et al. [6], and Maji et al. [7], who evaluated the performance of CNN-based CBIRS architectures against traditional computer vison methods, and concluded that the first performed better than the latter in terms of feature extraction, high-level pattern recognition, and creating highly discriminative embeddings where similar images are mapped into close proximity within a feature space, facilitating efficient retrieval through the comparison of feature vector distances. Naturally, with the recent advancement in transformers, similar investigations such as Denner et. al, [8], Song et al. [9] and Dubey et al. [10] published pioneer studies on ViT-based retrieval systems, showcasing their performance across multiple datasets. Similarly, the current study employed pre-trained ViTs for BrCa image retrieval, for evaluating the accuracy and efficiency in predicting post-surgery aesthetic outcomes compared with CNNs.

4 Dataset

The dataset comprises 2,193 instances, including clinical attributes (e.g., height, weight, age, bra size) and corresponding JPEG images of patients' upper torsos. For creating the ground-truth annotations, clinicians annotated 10-15 most similar images to a given query image such that the sample ranked i is more similar to the query image compared to the one appearing later in the list with rank $i+k$, where $k > 0$. By doing so, the dataset was split into 160 catalogues (80% for training, 20% for testing), with triplet-loss training using combinations of query, positive, and negative samples. Last, it is worth noting that the dataset includes both excellent and fair aesthetic outcomes for a given query, which are evaluated using different models.

5 Proposed Method

This study presents four techniques for patient retrieval, focusing on clinical data, image data, and a combination of both. The first approach uses clinical (tabular) data, employing weighted euclidean distance (WED) to optimize retrieval based on clinicians' rankings. Unlike standard Euclidean distance, WED uses a weight matrix to learn relationships between features. A shallow 4-layer multi-layer-perception (MLP) was also employed to better capture complex interactions between clinical features. The second experiment focuses on image data, utilizing pre-trained models including both CNNs and ViTs for feature extraction. These models were not fine-tuned in this phase and performed zero-shot feature extraction without updating model weights. In the third experiment, these same models were fine-tuned using triplet-loss to align the extracted image features with clinicians' rankings, ensuring semantic consistency with their objectives. The final experiment combines both modalities, concatenating fine-tuned image feature vectors with clinical data. The combined vectors were input into a shallow MLP, further refined using triplet-loss. This multimodal approach improves the retrieval of similar cases, enhancing the prediction of aesthetic outcomes. A brief overview of the described retrieval process is also illustrated in Figure 1.

[1] https://github.com/MsainZn/bcs-aesth-mmodal-retrieval

6 Results and Discussion

Clinicians typically require retrieval of both positive and negative treatment outcomes to illustrate diverse scenarios to patients. Following the same perspective, we also developed distinct retrieval models based on varying aesthetic assessments. This separation enables our algorithm to specifically address these preferences, thereby improving its ability to accurately reflect the needs of clinicians. The results, presented in Table 1, highlight the performance of different methods and models.

Tabular Data-Based Retrieval: The first approach used WED and a MLP for retrieval based solely on clinical data. While WED served as a baseline, MLP outperformed it but still achieved limited performance, with a maximum of 66% ADCG and 65% Acc. These results indicate that clinical data alone are insufficient for accurately predicting aesthetic outcomes due to their limited complexity.

Image-Based Retrieval: The second approach used both pre-trained and fine-tuned models for image retrieval. Pre-trained models generally performed worse than tabular models, possibly due to their misalignment with aesthetic criteria. However, fine-tuning significantly improved performance. For instance, VGG16's Acc increased from 52.10% to 69.26% and ADCG from 46.71% to 76.87% after fine-tuning. ViTs, especially BEIT, outperformed CNNs, with GoogleViT emerging as the best performer in 5 out of 8 metrics, demonstrating the effectiveness of ViTs for medical image retrieval.

Multimodal Retrieval: The final approach combined tabular and image features, yielding the best results with a 1-2% improvement in retrieval accuracy across all metrics. The multimodal method also improves robustness, as clinical data helps maintain reliability when models encounter out-of-distribution samples or underperform. This ensures consistent and informed decision-making in clinical settings, making the multimodal approach the most effective for aesthetic outcome prediction in BrCa treatment.

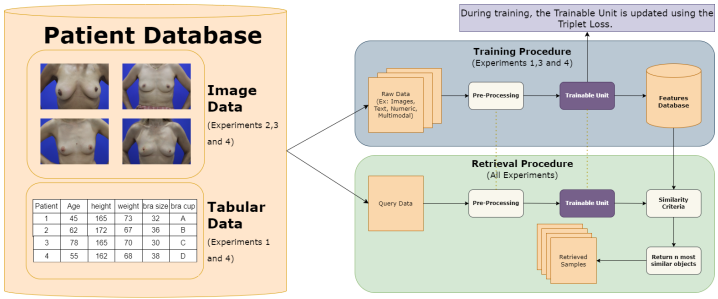

Figure 1: Illustration of the training and retrieval procedure. During training, the system optimizes the trainable unit using triplet loss to match clinicians' rankings. In retrieval, the system extracts features from a query input (image and text), and calculates similarity distance to find the most similar samples. Afterwards, image feature vectors can be stored in a feature database to accelerate process for future use.

7 Conclusions

In this study, triplet-loss optimization is applied to evaluate different clinical and image-based retrieval systems for predicting aesthetic outcomes in BrCa patients following surgical operations. Based on the results, clinical data alone showed moderate performance, highlighting its limitations in capturing complex aesthetic outcomes. Image models, especially ViT, outperformed traditional CNNs like VGG16 and ResNet. Finally, multimodal models, especially BEIT, performance aligned similar with clinicians' aesthetic criteria after fine-tuning. Future work can explore image-segmentation baselines to further enhance performance and address out-of-distribution challenges, given the dataset's limited variability.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389, and by FCT – Fundação para a Ciência e a Tecnologia within the PhD grants "2020.06434.BD" and "2024.06378.BD".

Table 1: Retrieval performance metrics for various models categorized by aesthetic quality. Each entry is of the form **(Excellent/Good, Fair/Poor)**.

Model	Train		Test	
	Acc (%)	ADCG (%)	Acc (%)	ADCG (%)
Tabular				
Baseline	(61.37, 61.77)	(65.21, 63.52)	(55.68, 61.87)	(57.75, 62.05)
Matrix	(60.41, 60.40)	(65.29, 62.02)	(**57.58**, 61.58)	(58.54, 65.31)
MLP	(**62.65**, **64.89**)	(**66.64**, **64.89**)	(57.09, **64.90**)	(57.70, **66.56**)
Pre-trained				
VGG16	(52.02, 46.68)	(54.55, 46.42)	(52.10, 44.72)	(56.69, 46.71)
ResNet	(54.96, 47.81)	(58.75, 50.56)	(53.38, 50.56)	(55.20, 53.74)
GoogleViT	(53.24, 47.29)	(56.21, **49.99**)	(52.10, 49.09)	(54.87, 53.13)
DINOv2	(56.00, 46.31)	(58.88, 47.82)	(**54.88**, 46.15)	(**60.231**, 44.95)
BEIT	(**56.03**, **49.25**)	(**61.06**, 49.39)	(54.79, **50.98**)	(57.58, **54.35**)
Fine-tuned				
VGG16	(90.51, 90.55)	(95.18, 95.42)	(69.26, 71.16)	(71.96, 76.87)
ResNet	(89.47, 87.11)	(94.22, 93.52)	(68.33, 72.17)	(70.30, 78.29)
GoogleViT	(**92.65**, 91.95)	(**96.45**, **96.28**)	(69.30, **73.18**)	(73.87, **79.86**)
DINOv2	(89.39, 87.77)	(94.38, 93.82)	(69.92, 70.78)	(**77.89**, 76.97)
BEIT	(89.78, **95.33**)	(93.95, 81.88)	(**71.95**, 72.76)	(73.85, 77.92)
Multimodal				
GoogleViT	(98.69, **98.91**)	(99.39, 99.56)	(68.55, **73.85**)	(71.14, **80.62**)
DINOv2	(97.89, 98.25)	(99.10, 99.33)	(70.14, 73.56)	(**77.40**, 78.96)
BEIT	(**99.10**, 98.87)	(**99.65**, **99.65**)	(**72.06**, 73.05)	(73.63, 78.45)

References

[1] Jaime S. Cardoso, Wilson Silva, and Maria J. Cardoso. "Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment". In: *The Breast* 49 (2020), pp. 123–130. ISSN: 0960-9776. DOI: https://doi.org/10.1016/j.breast.2019.11.006.

[2] Waqar M. Naqvi et al. "AI in Medical Education Curriculum: The Future of Healthcare Learning". In: *European Journal of Therapeutics* (2024). ISSN: 2564-7784. DOI: 10.58600/eurjther1995.

[3] Guangzhi Wang et al. "Semantic-Aware Triplet Loss for Image Classification". In: *IEEE Transactions on Multimedia* 25 (2023). ISSN: 19410077. DOI: 10.1109/TMM.2022.3177929.

[4] Shiv Ram Dubey. "A Decade Survey of Content Based Image Retrieval Using Deep Learning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32 (5 2022). ISSN: 15582205. DOI: 10.1109/TCSVT.2021.3080920.

[5] Huafeng Wang et al. "Deep Learning for Image Retrieval: What Works and What Doesn't". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 2015, pp. 1576–1583. DOI: 10.1109/ICDMW.2015.121.

[6] Vijayakumar Bhandi and K. A. Sumithra Devi. "Image Retrieval by Fusion of Features from Pre-trained Deep Convolution Neural Networks". In: *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*. 2019, pp. 35–40. DOI: 10.1109/ICATIECE45860.2019.9063814.

[7] Subhadip Maji and Smarajit Bose. "CBIR using features derived by Deep Learning". In: *CoRR* abs/2002.07877 (2020). arXiv: 2002.07877. URL: https://arxiv.org/abs/2002.07877.

[8] Stefan Denner et al. *Leveraging Foundation Models for Content-Based Medical Image Retrieval in Radiology*. 2024. arXiv: 2403.06567 [cs.CV].

[9] Chull Hwan Song et al. *Boosting vision transformers for image retrieval*. 2022. arXiv: 2210.11909 [cs.CV].

[10] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. "Vision Transformer Hashing for Image Retrieval". In: *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 2022, pp. 1–6. DOI: 10.1109/ICME52920.2022.9859900.