State of the Art: Computer Vision for Breast Tissue Analysis

Lina Lagzouli, Lamia Ladraa, Alexandre Morel, Grégoire Woroniak, Lucas Barrez
Computer Vision Course

December 2025

Abstract—**This state-of-the-art review examines recent advances in computer vision for breast cancer histopathology analysis. We focus on three complementary approaches: supervised learning with CNNs and Vision Transformers, zero-shot classification using vision-language models, and self-supervised learning for embedding analysis. Our review highlights the strengths and limitations of each paradigm, with particular attention to their application on the BreakHis dataset, and identifies gaps that motivate our multi-strategy approach combining interpretability and performance.**

I. Introduction

Breast cancer remains one of the most prevalent cancers worldwide, with early detection being crucial for successful treatment. Histopathological analysis of tissue samples is the gold standard for diagnosis, but access to expert pathologists is limited in many regions, particularly in medical deserts. Computer vision has emerged as a promising tool to assist in the screening and classification of breast tissue images.

Recent advances in deep learning have led to three main paradigms for medical image analysis: supervised learning with labeled data, zero-shot classification leveraging vision-language models, and self-supervised learning that extracts meaningful representations without explicit labels. Each approach offers unique advantages and faces specific challenges when applied to histopathology.

This review synthesizes the state of the art in these three areas, with a focus on breast cancer classification using the BreakHis dataset. We examine the performance, interpretability, and practical applicability of different methods, and identify research gaps that our project aims to address.

II. Dataset Presentation

As part of this project, we chose to use the BreakHis dataset, a widely used reference in research on the automatic diagnosis of breast cancer from histopathological images. This dataset was created by the P&D Laboratory and made publicly available in 2016 to promote the development and comparison of classification models in the biomedical field.

The dataset contains 9,109 microscopic images of breast tissue samples collected from 82 different patients. The samples were acquired using an optical microscope at four different magnification levels: ×40, ×100, ×200, and ×400. This diversity in resolution allows for the evaluation of a model's ability to generalize across varying levels of visual detail.

The images are divided into two main categories:

- **Benign tumors** (2,480 images), including four subtypes: Adenosis, Fibroadenoma, Phyllodes Tumor, and Tubular Adenoma;
- **Malignant tumors** (5,429 images), including four subtypes: Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma, and Papillary Carcinoma.

Each image is in RGB format (700×460 pixels) and was labeled by a pathologist to ensure reliable annotations. The dataset includes 8 classes and 4 magnification levels. It is worth noting that the dataset is imbalanced, with approximately 69% malignant images, a distribution that partly reflects real clinical conditions. This imbalance poses challenges for model training and requires careful evaluation strategies, particularly focusing on metrics like recall to avoid missing malignant cases.

III. Research Gaps and Project Positioning

Despite significant progress in applying computer vision to histopathology, several gaps remain. Most studies focus on a single learning paradigm (supervised, zero-shot, or self-supervised) without systematic comparison on the same dataset. This makes it difficult to assess the relative strengths of different approaches.

Interpretability remains a challenge. While high accuracies have been achieved, understanding which image regions drive predictions is crucial for clinical trust and validation. Attention maps and gradient-based visualization methods provide some interpretability, but their reliability and clinical relevance require further investigation.

Zero-shot and self-supervised methods have been less extensively evaluated on BreakHis compared to supervised approaches. Given the dataset's limited size and patient diversity, understanding how these methods perform without extensive labeled training data is valuable, particularly for scenarios with limited annotation resources.

Our project addresses these gaps through a multi-strategy approach that combines:

- Supervised learning with modern architectures (ResNet, EfficientNet) to establish strong baseline performance
- Zero-shot classification with CLIP to explore whether vision-language models can generalize to histopathology
- Embedding analysis with DINO to understand the intrinsic structure of tissue morphology
- Interpretability focus through probability matrices and heatmaps to identify prediction-relevant regions

This comprehensive evaluation will provide insights into which approach is most suitable for breast cancer screening in resource-constrained settings, where labeled data, computational resources, or expert supervision may be limited.

IV. Supervised Learning Approaches

A. Convolutional Neural Networks

CNNs have been the dominant architecture for histopathology image classification over the past decade. Transfer learning from ImageNet pre-trained models has proven particularly effective, as these models capture general visual features that transfer well to medical images despite domain differences.

Early work on BreakHis used classical architectures like AlexNet and VGG, achieving accuracies around 80-85% for binary classification. More recent architectures have substantially improved performance. Studies using ResNet variants have demonstrated strong results across different magnification levels, with deeper networks generally performing better on higher magnifications where more detailed features are visible.

EfficientNet models, which balance network depth, width, and resolution, have shown excellent performance on BreakHis. Research reports accuracy of 98.42% at ×400 magnification for binary classification, demonstrating the effectiveness of compound scaling strategies. DenseNet architectures, which use dense connections between layers, have also achieved competitive results while maintaining reasonable computational efficiency.

Recent work has explored ensemble approaches that combine multiple CNN architectures. One study combining DenseNet-201, ResNet-101, and NasNetMobile achieved 99.2% accuracy for binary classification by leveraging the complementary strengths of different architectures. However, these ensemble methods increase computational costs and may be less practical for deployment in resource-constrained settings.

B. Vision Transformers

Vision Transformers (ViT) represent a paradigm shift from convolutional architectures, using self-attention mechanisms to process images as sequences of patches. While initially requiring massive datasets, techniques like pre-training and data augmentation have made ViTs viable for medical imaging tasks.

Recent applications of ViT to BreakHis have demonstrated remarkable performance, with studies reporting accuracy up to 99.99% for binary classification. The self-attention mechanism allows ViTs to capture long-range dependencies in images, which may be beneficial for identifying distributed patterns in tissue morphology. However, ViTs typically require more training data and computational resources than CNNs, and their attention maps, while interpretable, can be more difficult to visualize meaningfully than CNN activation maps.

The trade-off between CNNs and ViTs involves computational efficiency, data requirements, and interpretability. For our project, exploring both architectures will provide insights into which approach best balances performance and practical constraints.

V. Zero-Shot Classification with Vision-Language Models

A. CLIP and Medical Imaging

CLIP (Contrastive Language-Image Pre-training) learns joint embeddings of images and text, enabling zero-shot classification by computing similarity between image embeddings and text descriptions of classes. While CLIP was trained on natural images, researchers have explored its application to medical imaging.

The key challenge is the domain gap between natural images and histopathology. Tissue morphology, color patterns, and spatial structures differ significantly from everyday objects. Studies show that vanilla CLIP performs poorly on medical images without adaptation. This has motivated research into domain-specific fine-tuning and prompt engineering.

CPLIP proposed adapting CLIP to histopathology through complete alignment of images and texts using pathological dictionaries and enriched textual descriptions. This approach improves the model's ability to distinguish between subtle morphological differences in tissue types.

B. Prompt Engineering Strategies

The effectiveness of zero-shot classification heavily depends on prompt design. Simple class names (e.g., ”fibroadenoma”) provide limited context, while detailed descriptions (e.g., ”histopathological image showing fibroadenoma with well-defined borders and epithelial proliferation”) improve performance by providing richer semantic information.

Research has explored hybrid approaches combining global and local image embeddings with prompt selection strategies. Some methods generate multiple prompts per class and select the best-performing ones based on validation performance. This prompt engineering process is crucial for maximizing zero-shot performance on specialized domains like histopathology.

For our project, we will experiment with different prompt formulations to assess their impact on classification accuracy and determine whether CLIP can meaningfully distinguish between histological subtypes despite being trained primarily on natural images.

VI. Self-Supervised Learning and Embedding Analysis

A. DINO for Histopathology

DINO (Self-Distillation with No Labels) is a self-supervised learning method that trains Vision Transformers to produce consistent representations across different views of the same image. Unlike contrastive methods, DINO does not require negative samples, making it simpler to implement and more memory-efficient.

Studies have shown that Vision Transformers trained with DINO learn interpretable visual concepts in histopathology and effectively localize cellular structures without supervision.

The learned representations capture hierarchical spatial information, with different attention heads focusing on different tissue components (nuclei, stroma, glandular structures).

DINO embeddings have demonstrated strong performance for downstream tasks like tissue classification and phenotyping. The quality of embeddings can be assessed through clustering analysis, with well-separated clusters indicating that the model has learned meaningful distinctions between tissue types.

B. Applications and Evaluation

Self-supervised embeddings enable several analysis approaches relevant to our project. K-nearest neighbor (k-NN) classification in embedding space provides a simple baseline that does not require training a separate classifier. Prototype-based methods compute class prototypes as the mean embedding of training samples, then classify new samples based on distance to prototypes.

Dimensionality reduction techniques like t-SNE and UMAP allow visualization of embedding spaces to assess whether histological subtypes naturally cluster. Good clustering indicates that the embedding space captures meaningful morphological differences. Recent work has reported silhouette coefficient improvements of 43% for tissue substructure segmentation using self-supervised representations.

Some research has integrated DINO with architectural modifications. CypherViT combined with DINO produces cleaner, less noisy clusters for tissue phenotyping, demonstrating that careful architecture design can enhance self-supervised learning outcomes.

For our project, we will use DINO embeddings to explore the intrinsic structure of BreakHis data and assess whether different histological subtypes form distinct clusters without supervised training. This will provide insights into the discriminative information present in the data and complement our supervised classification results.

VII. Evaluation Metrics in Medical Context

In medical diagnostics, false negatives (failing to detect cancer) are typically more costly than false positives (flagging healthy tissue for further examination). This asymmetry motivates the use of recall (sensitivity) as a primary metric, as it measures the proportion of actual positive cases correctly identified.

However, recall alone is insufficient for comprehensive evaluation. Precision measures the proportion of positive predictions that are correct, and the F1-score provides a harmonic mean of precision and recall. AUROC (Area Under the Receiver Operating Characteristic curve) evaluates performance across all classification thresholds and is particularly useful when class distributions are imbalanced.

Confusion matrices provide detailed information about which classes are confused with each other, which is valuable for understanding model failure modes. In multi-class settings, per-class metrics reveal whether the model performs consistently across all subtypes or struggles with specific categories.

For BreakHis, evaluation can occur at the image (patch) level or patient level. Patient-level evaluation aggregates predictions across multiple patches from the same patient, providing a more clinically relevant assessment. Our project will focus primarily on patch-level metrics but acknowledge the importance of patient-level aggregation for practical deployment.

VIII. Conclusion

This review has examined three complementary paradigms for breast cancer histopathology classification: supervised learning with CNNs and Vision Transformers, zero-shot classification with vision-language models like CLIP, and self-supervised learning using methods like DINO. Each approach offers unique advantages: supervised methods achieve high accuracy with sufficient labeled data, zero-shot methods potentially generalize to new classes without retraining, and self-supervised methods extract meaningful representations without labels.

The BreakHis dataset provides a valuable benchmark for evaluating these approaches, though its limited patient diversity and class imbalance present challenges. Recent work has achieved very high accuracies on this dataset using supervised methods, but questions remain about generalization, interpretability, and performance in low-data scenarios.

Our project will systematically compare these three paradigms on BreakHis, with particular emphasis on interpretability through attention visualization and probability mapping. By combining multiple approaches, we aim to provide a comprehensive assessment that can guide the development of practical tools for breast cancer screening in settings with limited specialist access.

References

[1] F. Spanhol et al., “A Dataset for Breast Cancer Histopathological Image Classification,” *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 7, pp. 1455-1462, 2016.

[2] F. Spanhol et al., “Deep Features for Breast Cancer Histopathological Image Classification,” *IEEE SMC*, 2017.

[3] M. Z. Alom et al., “Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network,” *Journal of Digital Imaging*, vol. 32, pp. 605-617, 2019.

[4] K. He et al., “Deep Residual Learning for Image Recognition,” *CVPR*, 2016.

[5] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *ICML*, 2019.

[6] “Explainable Soft Attentive EfficientNet for breast cancer classification in histopathological images,” *Biomedical Signal Processing and Control*, vol. 90, 2024.

[7] “Enhanced Histopathology Image Feature Extraction using EfficientNet with Dual Attention Mechanisms and CLAHE Preprocessing,” *arXiv:2410.22392*, 2024.

[8] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR*, 2021.

[9] “High-Performance Classification of Breast Cancer Histopathological Images Using Fine-Tuned Vision Transformers on the BreakHis Dataset,” *bioRxiv*, 2024.

[10] S. Javed et al., “CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment,” *CVPR*, 2024.

[11] “Path-CLIP: Bridging the Pathology Domain Gap - Efficiently Adapting CLIP for Pathology Image Analysis with Limited Labeled Data,” *ECCV*, 2024.

[12] Z. Zhao et al., “CLIP in Medical Imaging: A Comprehensive Survey,” *arXiv:2312.07353*, 2023.

[13] M. Caron et al., “Emerging Properties in Self-Supervised Vision Transformers,” *ICCV*, 2021.

[14] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, 2024.

[15] L. Ayzenberg et al., “DINOv2 Based Self Supervised Learning for Few Shot Medical Image Segmentation,” *ISBI*, 2024.

[16] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, “DINOv3,” *arXiv preprint arXiv:2508.10104*, 2025.