Journal of Pathology informatics

Research article

Empirical evaluation of deep learning-based feature extractors for object retrieval in digital pathology

ABSTRACT: This work investigates object image retrieval in digital pathology. We build upon models from recent literature to evaluate their performance across three large-scale object (patch) datasets, testing how well these models generalize to diverse types of pathology images. Our study provides an extensive comparison of 24 models under various conditions. Specifically, we examine the impact of the training dataset (in-domain vs. out-of-domain) and the training strategy (supervised vs. self-supervised). We also analyze how effectively foundation models perform when applied to unseen data. Supervised models yield the highest performance quantitatively when evaluated on data drawn from the same distribution as their training sets. Among the self-supervised models, although out-of-domain models generalize well to several types of data, in-domain foundation models exhibit superior performance overall and therefore should represent the preferred choice in most scenarios. Finally, we provide source code to run similar experiments and have integrated these tools into the Cytomine open-source web application to enable object search directly within its web user interfaces.

Keywords: Digital pathology, Patches, Image retrieval, Deep learning, Foundation models

1. Introduction

Histopathology is a complex field that plays a critical role in disease detection and diagnosis. Traditionally, these tasks require pathologists to examine large volumes of tissue samples under a microscope, a process that is time-consuming, costly, and highly dependent on specialized expertise. Digital and computational pathology aim to improve the efficiency and accuracy of this workflow. However, despite the success of deep learning in natural image analysis, histopathological images differ significantly from natural images, making the direct application of existing computer vision methods challenging.

Histopathological datasets exhibit substantial variability due to differences in tissue preparation and staining protocols, resulting in color distributions that differ markedly from those of natural images. In addition, tissue samples can be imaged at various magnifications, from detailed cellular views to broader tissue-level perspectives. Histopathology images are also often extremely large, particularly in the form of whole-slide images (WSIs), which must typically be divided into smaller patches before being processed by deep learning models. Furthermore, histopathological images are less amenable to standard image augmentations than natural images. For example, cellular morphology is critical for tumor detection, and excessive geometric distortions may alter clinically relevant features, introducing incorrect inductive biases and reducing model performance. These challenges highlight the need for methods specifically designed for histopathological data, particularly for applications such as disease classification [1,2], cell segmentation [3], survival / gene prediction [4,5], tissue phenotyping [6], and cancer prognosis [7].

Early computer vision research primarily focused on supervised learning. However, the abundance of unlabeled data relative to labeled data has driven a shift toward self-supervised learning (SSL), which has become a major research direction across many domains [8,9,10,11], including histopathology [12,13,14,15]. SSL enables models to learn useful representations from large collections of unlabeled data, reducing the need for costly and time-consuming annotation.

This capability is particularly valuable in histopathology, where the rapid digitalization of pathology workflows has generated vast amounts of uncurated data. Moreover, annotations are often available only at the slide level, providing limited information about the individual regions and structures within a slide [16,17,18]. As a result, SSL shows promise as an approach for leveraging these large datasets while potentially reducing the reliance on detailed expert labeling.

Alongside the growing interest in self-supervised learning, significant attention has recently been devoted to developing models that can generalize across diverse data types and applications. This effort has led to the emergence of foundation models, which aim to learn transferable representations that can be applied to multiple downstream tasks and domains without requiring full model retraining. In recent years, several vision foundation models have been introduced for both natural image analysis, including DINO [8], CLIP [19], and iBOT [20], and histopathology, such as UNI [21], Phikon [22], and H-optimus [23]. These models have become widely adopted as feature extractors and backbones in subsequent research.

Image retrieval is a computer vision task that aims to identify and retrieve, from a database, the images whose content is most similar to a given query image. In digital pathology, such methods can facilitate the identification of similar, potentially rare, cells or tissue regions within large image repositories. This capability can support diagnostic workflows by retrieving previously diagnosed cases containing similar patterns, as well as annotation tasks by assigning labels based on the majority class among the retrieved samples. Additional applications of content-based image retrieval (CBIR) in digital pathology are discussed in Section 2.2.

Image retrieval frameworks typically comprise two main components: feature extraction and search. The feature extraction component generates representative embeddings from images, while the search component indexes these embeddings and retrieves the most relevant matches from a database. Retrieval is commonly based on nearest-neighbor methods combined with similarity metrics such as Euclidean distance or cosine similarity. In modern content-based image retrieval (CBIR) systems, feature extraction relies on deep learning models and has become the primary focus of research. Numerous architectures and training strategies have been proposed to improve the quality and transferability of the learned representations. This work focuses specifically on the feature extraction component, as the quality of the extracted features is a key determinant of the overall retrieval performance.

More specifically, this work provides a comprehensive evaluation of feature extraction methods for patch-level object retrieval in digital pathology. Our main contributions are as follows:

a. We evaluate state-of-the-art models trained using supervised learning and self-supervised learning (SSL). To ensure broad coverage of SSL paradigms, we select representative models from the major SSL method families in computer vision, namely DINO [8,80], BYOL [10], and iBOT [20].
b. We investigate the impact of training data on retrieval performance by comparing models trained on in-domain (histopathology) and out-of-domain (natural image) datasets. We also evaluate performance across multiple histopathology datasets to assess generalization within the domain.
c. We examine the effectiveness of recent histopathology foundation models when used directly for retrieval without task-specific fine-tuning. The evaluated models include UNI [21], RetCCL [13], Phikon [22], cdpath21-gan [24], CTransPath [25], Virchow [26], and H-optimus [23].
d. All codes developed in this work are released under a permissive open-source license. We provide a Python toolbox for reproducing our experiments, as well as an integrated retrieval system within the open-source Cytomine platform [73], incorporating one of the best-performing feature extractors identified in our study.

2. Related works

Image retrieval relies on learning informative representations of input data, making feature extraction a fundamental component of retrieval systems. We therefore begin by reviewing the state of the art in feature extraction methods (Section 2.1). Next, we present approaches specifically developed for image retrieval (Section 2.2). Finally, we discuss studies in the histopathology domain that are most closely related to our work (Section 2.3).

2.1. Feature Extraction

Image retrieval relies on feature extraction to generate representations that enable image comparison. Representing images as feature vectors has long been a fundamental aspect of computer vision, evolving from handcrafted descriptors in traditional machine learning approaches to deep-learning-based representations in modern systems. The quality of these representations is critical for downstream tasks, including content-based image retrieval (CBIR), and depends on the ability to capture image characteristics relevant to the target application. This section reviews both supervised and self-supervised feature learning approaches before discussing foundation models for natural images and digital pathology.

Supervised Approaches

Early deep-learning-based feature extraction methods were primarily derived from supervised models trained on large annotated datasets such as ImageNet [27]. Although originally developed for image classification, these models have since been widely adopted for a variety of computer vision tasks [28,29]. Convolutional neural networks (CNNs) have long been the dominant architecture for learning image representations [30]. For example, ResNet-based models have been used for image retrieval [31,32] and face recognition [33]. More recently, Vision Transformers (ViTs) have emerged as a powerful alternative for feature extraction [34]. Both CNNs and transformer-based architectures have been successfully applied to a wide range of histopathology tasks. For instance, ResNet backbones have been employed for whole-slide image classification [35], while transformer-based models combined with multiple-instance learning have also demonstrated strong performance on the same task [36].

Self-Supervised Approaches

Self-supervised learning (SSL) differs from supervised learning in that feature representations are learned without relying on labeled data. Most SSL methods aim to learn feature spaces in which similar images are represented by nearby embeddings, while dissimilar images are mapped farther apart. Since labels are unavailable, SSL approaches differ primarily in how they define similarity and dissimilarity between samples. Existing SSL methods can be broadly categorized into four families: contrastive learning (e.g., SimCLR [37]), non-contrastive learning (e.g., BYOL [10]), clustering-based methods (e.g., SwAV [38]), and generative learning (e.g., cdpath21-gan [24]). A more detailed overview of these categories is provided in Supplementary Material A.

Foundation Models for Natural Images and Digital Pathology

Foundation models have attracted considerable attention for their ability to learn representations that transfer effectively across a wide range of tasks and datasets. These models have demonstrated strong generalization capabilities and have become widely adopted in computer vision. In particular, DINO [8] and CLIP [19] have emerged as influential models and training frameworks, inspiring numerous adaptations and extensions across different domains. DINO, which builds on principles introduced by BYOL, is described in more detail in Section 3.1.1. While early foundation models were primarily developed for natural images, recent years have seen the emergence of several foundation models specifically designed for histopathology. Notable examples include UNI [21], RetCCL [13], Phikon [22], cdpath21-gan [24], CTransPath [25], Virchow [26], and H-optimus [23]. As these models are included in our study, they are described in detail in Section 3.1.1.

2.2. Image Retrieval Systems and Applications in Digital Pathology

Several studies have developed feature extraction methods specifically for image retrieval in supervised settings. Among these, Deep Metric Learning (DML) is one of the most widely used approaches [79]. DML learns representations in which images sharing the same label are mapped close to each other in the feature space, while images from different classes are pushed farther apart. In self-supervised learning (SSL), image retrieval is often used as a downstream task to evaluate the quality of learned representations. SSL-based retrieval performance typically depends on both the training strategy and the backbone architecture. For example, DINO [8] was originally evaluated on image retrieval tasks using a Vision Transformer (ViT) backbone. Some SSL methods have also been designed specifically for retrieval. For instance, InsCLR [43] introduces a self-supervised training strategy tailored to instance retrieval.

Several CBIR frameworks have been specifically tailored for histopathology, addressing the unique challenges of patch or whole slide image retrieval. SMILY [44] uses deep feature embeddings to retrieve similar regions across tissue types and diseases, while Yottixel [45] is a search engine designed to support pathology diagnosis. Both methods employ a mosaic-based strategy that clusters WSIs into representative patches. To enable efficient large-scale retrieval, they use compressed image representations, and SISH [46] further improves scalability through tree-based indexing. More recently, RetCCL [13] demonstrated that clustering-guided contrastive learning can produce transferable feature representations for multiple downstream tasks without additional fine-tuning.

The practical applications of CBIR in digital pathology extend across several clinical tasks. For diagnosis support, systems such as SMILY [44] and Luigi-Oral [48] retrieve cases with similar histological features or disease severity to help classify lesions as benign or malignant. Beyond direct diagnosis, CBIR has proven valuable for quality control, for example, it has been used to track the source of "floating" tissue contaminants by matching them to their slide of origin [47]. Additionally, retrieval-based methods offer a robust alternative when labeled data is too scarce for supervised learning, helping pathologists navigate large databases to find relevant reference information [46]. Even in specialized tasks like nuclei segmentation, CBIR-based approaches have been successfully integrated with traditional computer vision techniques to identify and delineate structures based on a reference template [49].

This diversity of applications highlights the need for rigorous empirical evaluations of CBIR frameworks. As foundation models continue to emerge, such studies can help identify the most effective architectures and learning strategies across diverse image types. They also provide a practical means of narrowing the large number of available models before conducting costly and time-consuming reader studies involving pathologists.

2.3. Empirical Studies in Histopathology

Several empirical studies have compared feature extraction methods in digital pathology. In this section, we review these studies and position our work relative to them.

Several empirical studies have compared feature extraction methods in digital pathology. For example, [50] evaluated 14 pathology-specific self-supervised models for WSI classification and found that the choice of feature extractor had a much greater impact than stain normalization or data augmentation, with LUNIT-DINO [14], UNI, and CTransPath achieving the best results. Similarly, [51] compared RetCCL, Yottixel, SISH, and HSHR [52] for patch and slide retrieval, identifying RetCCL as the strongest performer and emphasizing the importance of replicability, robustness, and clinical relevance when proposing new CBIR frameworks. Finally, [14] evaluated MoCo v2, Barlow Twins, SwAV, and DINO on histopathology tasks, showing that in-domain pretraining is beneficial but that no SSL method consistently outperforms the others. These findings are consistent with our observations (Section 4).

In [53], feature extractors are evaluated on clinically relevant weakly supervised tasks, including both vision-only and vision-language models. The authors show that the vision-language model CONCH consistently outperforms vision-only approaches. Among the latter, UNI achieves the best performance, in line with our image retrieval results. CTransPath and Phikon are also similarly ranked, which is consistent with our findings. In addition, the study highlights that dataset diversity has a greater impact on performance than dataset size.

Beyond comparative evaluations, some studies focus on the extensive assessment of a single model across multiple tasks and settings, as shown for Phikon [22] and UNI [21].

Our study differs from previous work in several respects. We evaluate both supervised and self-supervised feature extractors, including models trained on in-domain and out-of-domain data, whereas most studies focus on a single

setting. We also specifically investigate object retrieval in histopathology, a less explored task than whole-slide image classification. Finally, our datasets enable a more detailed analysis of model performance across diverse tissue and organ types, as well as fine-grained tissue structure.

3. Methods

This section presents the models and methods used in our experiments. Section 3.1 introduces the feature extraction, indexing, and retrieval methods evaluated in this study, while technical details are deferred to Supplementary Material B. Section 3.2 describes the training and benchmark datasets, Section 3.3 outlines the evaluation protocols, and Section 3.4 presents the implementation.

Name	Arch.	Training Data	LS Size	Training method	Embedding vector size	License
Supervised learning						
ResNet	ResNet50	ImageNet-1K+ULiege	633k P.	DML	128	/
DeiT	DeiT	ImageNet-1K+ULiege	633k P.	DML	128	/
SSL out-of-domain						
Dino pre-trained	ViT-S/16	ImageNet-1K	/	Dino	384	Apache 2.0
Dino fine-tuned	ViT-S/16	ImageNet-1K+ULiege	633k P.	Dino	384	Apache 2.0
Dino scratch	ViT-S/16	ULiege	633k P.	Dino	384	Apache 2.0
iBOT pre-trained	ViT-S/16	ImageNet-1K	/	iBOT	384	Apache 2.0
iBOT fine-tuned	ViT-S/16	ImageNet-1K+ULiege	633k P.	iBOT	384	Apache 2.0
iBOT scratch	ViT-S/16	ULiege	633k P.	iBOT	384	Apache 2.0
BYOL pre-trained	ResNet50	ImageNet-1K	/	BYOL	256	MIT
BYOL finetuned	ResNet50	ImageNet-1K+ULiege	633k P.	BYOL	256	MIT
BYOL scratch	ResNet50	ULiege	633k P.	BYOL	256	MIT
Dinov3 pre-trained	ViT-L/16	LVD-1689M	/	Dinov3	1024	custom
Dinov3 fine-tuned	ViT-L/16	LVD-1689M+ULiege	633k P.	Dinov3	1024	custom
Dinov3 scratch	ViT-L/16	ULiege	633k P.	Dinov3	1024	custom
SSL in-domain						
RetCCL	ResNet50	TCGA [17]/PAIP [54]	15M P.	Contr. & Clust.	2048	GPL-3.0
UNI	ViT-L/16	Mass-100k	100k W.	Dinov2	1024	Non-comm.
UNI v2	ViT-H/15	Proprietary	350k W.	Dinov2	1536	Non-comm.
CTransPath	Swin	TCGA	15M P.	Clustering	768	GPL-3.0
cdpath21-gan	ResNet18	Camelyon17 [18]	140k P.	Generative	512	/
Phikon	ViT-B/16	TCGA	40M P.	iBOT	768	Non-comm.
Phikon v2	ViT-L/14	PANCAN-XL	60k W.	Dinov2	1024	Non-comm.
H-optimus-0	ViT g14+reg.	Proprietary	500k W.	iBOT and Dinov2	1536	Apache 2.0
H-optimus-1	ViT-H/14+reg.	Proprietary	1M W.	Undocumented	1536	cc-by-nc-nd-4.0
Virchow v2	ViT-H/14	MSKCC	3.1M W.	Modified Dinov2	1280	cc-by-nc-nd-4.0

Table 1: The different models investigated in this work with their main characteristics: from left to right: the name given to the model, model architecture, training dataset, the size of the training dataset (P. means patches, W. means whole-slide image), the SL or SSL strategy used to train the model, the size of the embedding vector, and the license, if any, under which the model is available.

3.1. Retrieval Model

3.1.1 Feature Extraction

Table 1 summarizes the feature extraction models evaluated in this study. Models were selected based on their popularity and effectiveness in the literature, and to cover a broad range of training paradigms. Our selection includes Deep Metric Learning (DML) and contrastive, non-contrastive, generative, and clustering-based self-supervised learning. We first describe the supervised models, followed by the SSL approaches.

Supervised Methods

Supervised feature extractors are obtained by training neural network backbones on labeled datasets using a metric-learning objective that encourages images from the same class to have similar representations. We consider two backbones: ResNet50 [28], a widely used convolutional neural network, and DeiT [55], a Vision Transformer architecture. Both models are trained using Deep Metric Learning [56] (see Supplementary Material B.1).

Self-Supervised Methods

The SSL-based feature extractors considered in this study differ in their training objectives, the nature of their pre-training data, and their architectural design.

Natural Image-based Frameworks: We evaluate feature extractors trained with BYOL [10], DINO [8], DINOv3 [80], and iBOT [20]. Originally developed for natural-image datasets such as ImageNet-1K [27], these methods learn representations without labels using non-contrastive objectives. They differ in the mechanism used to prevent representation collapse (i.e., all images converging to the same representation). **BYOL** relies on a student-teacher architecture using a Mean Squared Error (MSE) loss to match representations between two augmented views of the same image. **Dino** also builds upon the student-teacher paradigm but introduces a distillation-based approach using centering and sharpening operations, rather than the predictor head used in BYOL. Furthermore, while BYOL relies on two global views, Dino employs a multi-crop strategy (typically eight views). By requiring the student to match the teacher's global features using only local patch information, Dino encourages the model to learn more robust semantic spatial features. **iBOT** further extends this idea by incorporating a masked image modeling (MIM) objective, treating the vision task similarly to masked language modeling, allowing finer local structures to be captured.

In our experiments, we use Dino and iBOT to train a ViT-S/16 architecture, BYOL to train a ResNet50 CNN backbone, and Dinov3 to train a ViT-L/16 transformer. These choices correspond to the most common defaults defined in the relevant publications, and for which configuration files were available. The choice for Dinov3's architecture to be a ViT-L, on top of being the default configuration in their codebase, also reflects the current tendency of the literature to turn to larger models, with more parameters.

For convenience, and since a single backbone architecture is trained for each SSL strategy, we will refer to the resulting feature extractor models as Dino, Dinov3, iBOT and BYOL, i.e., using the name of SSL training strategy from which each model is trained.

Histopathology-specific Foundation Models: We include seven domain-specific models: RetCCL [13], UNI [21], CTransPath [25], cdpath21-gan [24], Phikon [22], H-optimus [23], and Virchow [57]. These models vary in their backbone architectures (e.g., ResNet, Swin Transformer, or ViT), the SSL strategies used to train them (contrastive, non-contrastive, clustering-based, or generative), and the scale of their training datasets. For UNI, Phikon, and H-optimus, we evaluate two versions, allowing us to assess the impact of model scaling and larger pretraining datasets. Notably, while the original Phikon model was trained with iBOT, Phikon v2 adopts a DINOv2-based training strategy designed to improve feature consistency across global and local scales.

3.1.2 Similarity & Search

After feature extraction, image retrieval is performed by indexing the resulting feature vectors and retrieving the k nearest neighbors of a query vector according to a predefined distance measure.

To achieve this, we combine two open-source libraries: FAISS [58] and Redis. FAISS provides efficient similarity search algorithms and supports several distance metrics. In this work, we use exact (brute-force) nearest-neighbor search with Euclidean distance. Given a query vector, the system computes its distance to every indexed vector, ranks the results, and returns the k closest images. Although faster search strategies are available in FAISS, our focus is on feature extraction rather than retrieval efficiency. Since FAISS only supports integer identifiers for indexed vectors, Redis is used as an in-memory key-value store to map these identifiers to the corresponding image filenames.

3.2. Datasets

We evaluate our methods on three histopathology patch datasets. The first and largest dataset is used for both fine-tuning natural-image models and evaluation, while the two others serve as public benchmarks to assess out-of-distribution generalization.

ULiege dataset. Compiled from [59,60], this dataset contains approximately one million patches extracted from whole-slide images. It combines public and private datasets into a single collection spanning multiple organs, pathologies, staining protocols (e.g., H&E, IHC), and image resolutions, and combines data from 19 projects. Each project concerns different organs / tasks. For example, *Warwick crc* contains patches of the colon and focuses on the classification of nuclei, *Janowczyk* is a dataset related to breast cancer, and *Glomeruli* is for the kidneys. Each project defines its own classes, resulting in 67 classes overall, with 2 to 9 classes per project. The dataset is highly heterogeneous and class-imbalanced, with varying patch dimensions (Table C.1). For consistency, all patches were resized to 224×224. Table 2 summarizes the projects, while Figure 1 illustrates one patch per class.

The dataset was split into a training set (633k patches), an indexing set (106k patches), and a query set (96k patches). The training set is used for supervised training and for adapting SSL models pretrained on out-of-domain data. The indexing and query sets are used for evaluation. Splits were created to prevent data leakage by ensuring that all patches from a given patient or slide belong to the same subset, while maintaining class representation across all splits.

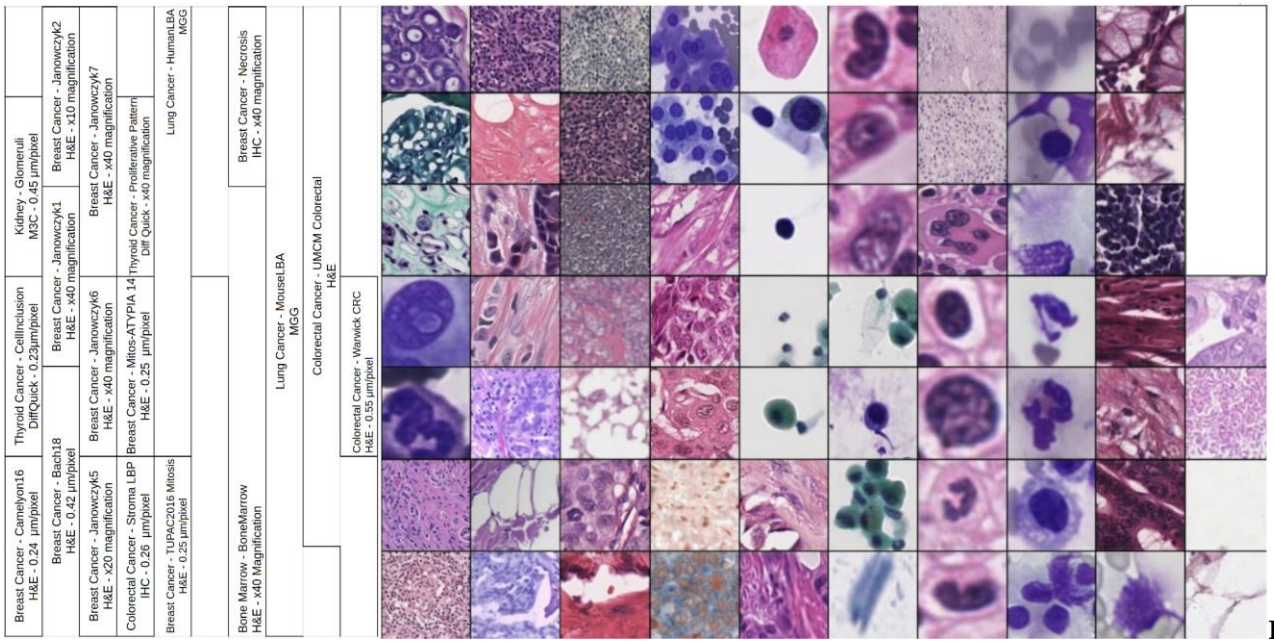

Figure 1: 67 patches, one for each class of the ULiege dataset. In the left part of the picture, the pathology, dataset name, stain and resolution are provided

Project	#c	#i	Project	#c	#i
Camelyon16 [61]	2	292,226	Ulg lbtd2 chimio necrose [59]	2	882
Cells no aug [59]	2	3,637	Ulg lbtd lba [59]	8	4,284
Patterns no aug [59]	2	1,857	Ulg bonemarrow [62]	8	1,291
Glomeruli no aug [63]	2	29,213	Tupac mitosis [64]	2	77,853
Iciar18 micro [65]	4	4,800	Warwick crc [66]	2	2,500
Janowczyk1 [67]	2	31,725	Umcm colorectal [68]	8	5,000
Janowczky2 [67]	2	3,402	Mitos2014 [69]	3	64,873
Janowczyk5 [67]	2	24,870	Ulb anapth lba [59]	9	5,420
Janowczyk6 [67]	2	277,456	Lbpstroma [69]	2	2,313
Janowczyk7 [67]	3	2,244	**Total**	**67**	**835,846**

Table 2: Split of the ULiege dataset (#c = number of classes, #i = number of images)

NCT-CRC-HE-100K [68]. This dataset contains 100k H&E-stained colorectal tissue patches (224×224) distributed across nine classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. Figure 2 shows sample patches from each class. We use NCT-CRC-HE-100K as the indexing set (100k patches) and CRC-VAL-HE-7K as the query set (7,180 patches). This dataset is widely used as a benchmark in the literature.

Camelyon17-WILDS [71]. Camelyon17 is a breast cancer lymph node dataset composed of H&E-stained slides [18]. We use Camelyon17-WILDS, its patch-based version from the WILDS benchmark collection. The dataset contains 96×96 patches belonging to two classes. Example patches are shown in Figure 3. For evaluation, the original OOD test set is used as the indexing set (85k patches), while the OOD validation set serves as the query set (33.5k patches).

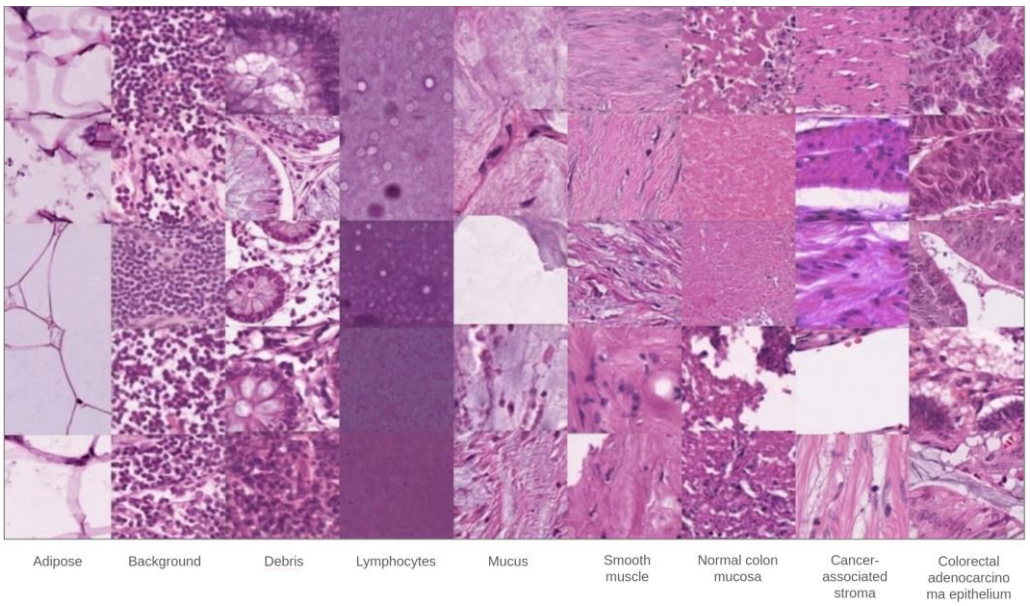

Figure 2: 5 patches of each class of the NCT-CRC-HE-100K dataset (one class per column) at the 0.5 μm/px resolution

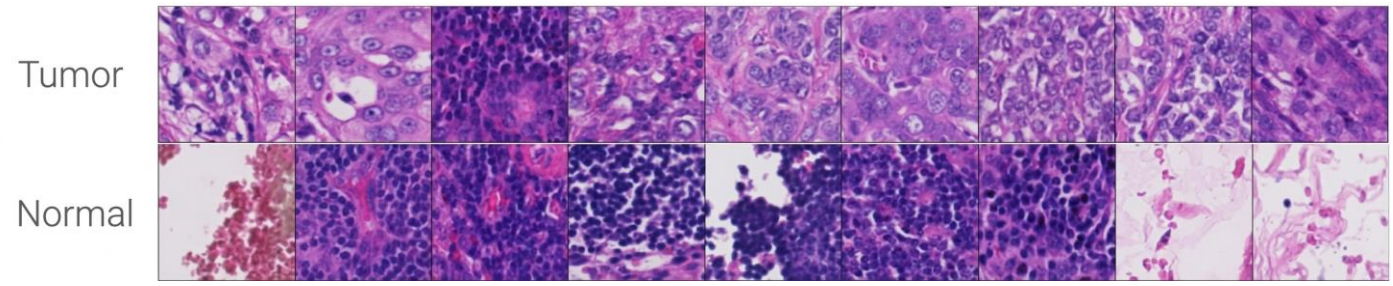

Figure 3: Patches of each class of the Camelyon17-WILDS dataset (one class per row) at 0.23µm–0.25µm/px resolution

3.3. Evaluation and Protocols

Evaluating CBIR systems is challenging because no universal definition of patch similarity exists. Following prior work, we assess retrieval performance quantitatively using image labels by comparing the query label with those of the retrieved patches. We also provide a qualitative evaluation in Appendix D. However, this analysis is inherently subjective, as it is based on a limited set of randomly selected queries.

3.3.1 Performance metrics

Retrieval performance is evaluated using Top-1, Top-5, and Maj-5 accuracy [21,13]. Top-k accuracy measures the proportion of queries for which at least one of the top k retrieved images belongs to the same class as the query. Maj-5 accuracy measures the proportion of queries for which at least three of the top five retrieved images share the query label. We additionally report precision, i.e., the proportion of retrieved images of the correct class, as a function of the number of retrieved images per query.

3.3.2 Protocols

For each model, the indexing set is encoded, and its feature vectors are stored in FAISS and Redis (Section 3.1.2). Query patches are then encoded individually, and FAISS retrieves the five nearest neighbors from the indexing set using Euclidean distance. Retrieval metrics are computed by comparing the query label with the labels of the retrieved patches. Results correspond to the average performance across all queries.

3.3.3. Training Details

For the **in-domain SSL foundation models** (e.g., UNI and Phikon), we use the official architectures and pretrained weights released by the authors.

The supervised models (ResNet50 and DeiT) are initialized with ImageNet-1K weights and trained on the ULiege dataset for 100 epochs using Adam (learning rate 10^{-4}) with weight decay. An exponential learning-rate scheduler (γ=0.3) is applied. To address class imbalance, training batches are sampled to ensure equal class representation.

The out-of-domain SSL frameworks (DINO, iBOT, and BYOL) are evaluated in three settings:

- **Pretrained:** ImageNet-1K weights used without additional training.
- **From Scratch:** Random initialization followed by 100 epochs of training on ULiege.
- **Fine-tuned:** ImageNet-1K initialization followed by 100 epochs of training on ULiege.

For the **From Scratch** and **Fine-tuned** settings, we follow the original training procedures and loss functions described in respective publications [8, 10, 20], with all network weights left unfrozen. Training hyperparameters are identical to those of the original publications.

For DINOv3, the **From Scratch** and **Fine-tuned** models are trained using the official GitHub implementation ("Training" section of the ReadMe) and the provided configuration file (*vitl_im1k_lin834.yaml*). Only the dataset configuration and batch size are modified to accommodate the ULiege dataset and available computational resources.

3.4 Source Code and Retrieval System Implementation

Our experiments were conducted in Python using our open-source library, which provides command-line scripts for reproducing object retrieval experiments (https://anonymous.4open.science/r/Empirical-Evaluation-of-Deep-Learning-based-feature-extractors). Beyond reproducing our results, the library facilitates the systematic comparison of retrieval approaches for digital pathology applications, helping researchers identify promising models before conducting more targeted expert or clinical evaluations (Section 7.3).

We also developed a user-friendly CBIR module for the open-source Cytomine web platform [73]. Its source code is also released under a permissive open-source license and integrated to the official Cytomine repository (https://github.com/cytomine). The module enables annotation similarity search directly from the Cytomine web interface and suggests labels based on those assigned to the ten most similar retrieved annotations. All user-generated annotations are indexed, allowing on-demand retrieval of similar annotations (Figure 4). The implementation supports multiple feature extractors, including ResNet50 and H-optimus-0. We selected H-optimus-0 by default because it achieved performance comparable to the best models in our experiments while being distributed under a permissive open-source license, unlike several competing approaches (Section 5.2).

The prototype is implemented as a Docker-based service using FastAPI, a high-performance framework for REST APIs. It listens to requests from the Cytomine web interface. When a new annotation is created, the corresponding image region is extracted, indexed, and sent to the retrieval service, which computes a feature vector and stores it in a FAISS index running within the Docker container.

For similarity search, the query annotation is encoded and compared against the indexed feature vectors using FAISS. The system retrieves the n nearest neighbors, maps them back to their Cytomine annotation identifiers, and displays the results in the Cytomine interface (Figure 4).

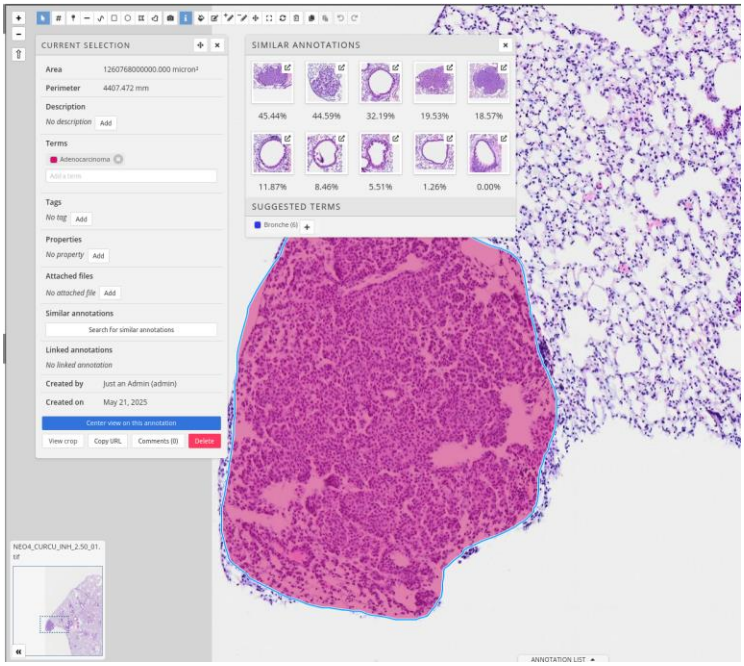

Figure 4: Screenshot of the implementation of the retrieval process in the open-source Cytomine software. Once an annotation geometry is drawn or selected, the user can click on the "Search similar annotations" button to view the ten most similar annotations according to the model H-optimus-0.

4. Results

The quantitative performance of all 24 models across the three datasets is summarized in Table 3, while Figure 5 shows the proportion of correctly retrieved images as a function of the number of retrieved neighbors. Additional qualitative results are provided in Supplementary Material D. These include t-SNE visualizations of the embeddings learned on the ULiege dataset (Figure D2) and examples of query images with their top-ranked retrievals across all models for selected classes from the three datasets (Figures D3-D8). As these analyses are based on a limited number of samples, they should be viewed as illustrative complements to the quantitative evaluation rather than as definitive evidence of model performance.

	ULiege dataset (all) (%)			NCT-CRC-HE-100K (%)			Camelyon17-WILDS (%)		
	Top-1	Top-5	Maj-5	Top-1	Top-5	Maj-5	Top-1	Top-5	Maj-5
Supervised									
ResNet	81.83	92.92	84.29	83.58	93.54	84.99	76.78	89.43	78.54
DeiT	79.77	91.95	82.01	80.67	91.64	81.73	68.61	86.60	70.97
SSL Out-of-Domain									
Dino pre-trained	75.91	90.38	77.02	93.59	97.79	93.83	70.32	85.44	72.06
Dino fine-tuned	76.87	88.98	77.75	94.39	98.01	94.89	76.91	88.54	79.10
Dino scratch	77.40	89.35	78.36	95.40	98.34	96.04	78.02	89.25	79.81
iBOT pre-trained	75.78	90.51	77.05	92.35	97.77	92.98	70.29	87.16	71.60
iBOT fine-tuned	77.74	89.50	78.78	94.92	98.41	95.93	81.39	93.00	83.34
iBOT scratch	76.65	88.89	77.64	94.28	97.83	95.14	79.56	92.29	82.49
BYOL pre-trained	63.80	84.32	62.60	70.93	90.29	70.43	66.36	88.65	69.50
BYOL fine-tuned	75.41	89.42	77.17	79.86	93.06	80.81	63.55	89.96	66.58
BYOL scratch	73.91	90.29	75.57	81.13	93.12	81.56	76.89	91.13	77.90
Dinov3 pre-trained	72.97	89.48	73.64	89.57	97.3	91.09	78.74	94.21	81.94
Dinov3 fine-tuned	73.61	91.7	74.54	76.62	89.21	77.28	63.83	81.32	66.27
Dinov3 scratch	73.67	90.37	75.16	82.3	94.51	83.61	39.75	63.86	42.29
SSL In-Domain									
Cdpath21-gan	58.39	78.75	57.01	52.86	77.77	48.83	73.00	91.60	75.96
CTransPath	76.42	89.44	77.41	96.07	98.44	96.31	80.16	92.01	82.50
RetCCL	73.68	87.64	73.85	91.41	97.38	92.30	79.04	89.59	79.80
UNI v1	79.34	90.74	80.78	95.99	97.26	96.46	93.89	97.90	95.50
UNI v2	79.61	91.39	80.82	96.88	96.57	94.68	92.54	97.40	93.38
Phikon v1	78.05	89.36	78.93	95.04	97.60	95.74	80.91	91.87	83.82
Phikon v2	78.20	90.00	79.34	94.87	97.02	95.29	73.91	86.20	76.13

H-optimus-0	75.80	89.17	78.44	94.22	96.57	94.68	92.67	96.76	93.38
H-optimus-1	77.80	90.23	78.93	95.35	97.41	96.17	92.61	98.64	96.07
Virchow v2	77.65	89.56	78.59	95.72	96.56	96.02	94.83	97.41	95.44

Table 3: Top-1, Top-5, and Maj-5 accuracy for the supervised and self-supervised models, both in-domain and out-of-domain, on the ULiege dataset, the NCT-CRC-HE-100K dataset and the Camelyon17-WILDS dataset. For each dataset and evaluation metric, the best-performing value is highlighted in color within each model family (supervised in blue, SSL out-of-domain in orange, and SSL in-domain in green).

Consistency across models. Top-1 and Maj-5 accuracies remain highly consistent across all models and datasets. This suggests that the models produce robust rankings, as their ability to retrieve a single correct image is similar to their ability to retrieve a majority of relevant images among the top results. This trend is also visible in Figure 5. For all models, performance remains nearly constant as retrieval depth increases. Although a slight decrease can be observed, it remains below 0.2%, indicating that the rankings are coherent and stable.

Top-5 accuracy further supports this observation. Compared with Top-1 accuracy, it improves by roughly 10% on the ULiege and Camelyon17-WILDS datasets and by 3 to 4% on NCT-CRC-HE-100K. This indicates that, even when the top-ranked result does not belong to the correct class, at least one of the next four retrieved images is likely to match the query.

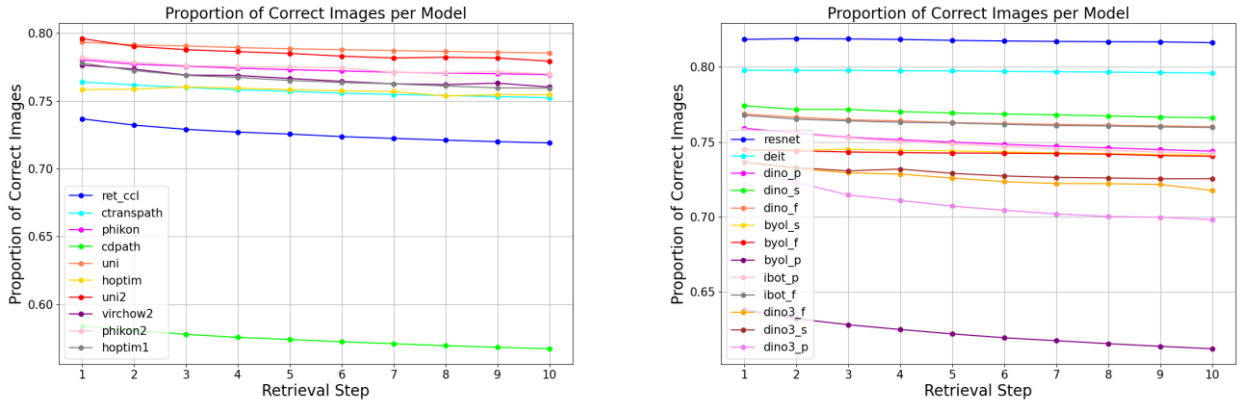

Figure 5: Proportions of images belonging to the same class as the query, for a range of retrieved images from 1 to 10, for the ULiege dataset. The supervised models and the out-of-domain SSL models are shown in the left plot, while the in-domain SSL models are shown in the right plot.

Supervised learning (versus SSL). The two supervised models (the first two rows of Table 3) achieve similar performance across all metrics, although ResNet consistently outperforms DeiT. Their performance on ULiege, the training dataset, is comparable to that on NCT-CRC-HE-100K and higher than on Camelyon17-WILDS. Supervised models also outperform self-supervised models on ULiege. However, this trend reverses on NCT-CRC-HE-100K and Camelyon17-WILDS, where SSL models achieve higher overall performance. These findings suggest that supervised learning remains advantageous in in-distribution settings, whereas self-supervised approaches generalize better to datasets with different distributions or limited label availability.

Out-of-domain SSL strategies. No self-supervised learning strategy consistently outperforms the others. BYOL generally ranks below iBOT and DINO across datasets, which is consistent with the evolution of these methods, as DINO was designed to address some limitations of BYOL. However, this comparison should be interpreted cautiously since BYOL uses a ResNet backbone, whereas DINO and iBOT rely on Vision Transformers. Notably, BYOL performs worse than the supervised ResNet trained with DML on ULiege, but achieves comparable results on the two benchmark

datasets. Despite being the most recent framework and using a larger backbone, DINOv3 yields mixed results. Its pretrained version outperforms BYOL overall and achieves the best performance on Camelyon17-WILDS, but remains below DINO and iBOT on ULiege and NCT-CRC-HE-100K.

The results nevertheless confirm the strong generalization ability of out-of-domain SSL models. Pretrained DINO and iBOT achieve Top-1 accuracies of about 75% on ULiege, 93% on NCT-CRC-HE-100K, and 70% on Camelyon17-WILDS. All three SSL frameworks, DINO, iBOT, and BYOL, generally benefit from additional training on histopathology data. Both fine-tuned and from-scratch variants improve performance not only on ULiege but also on the two benchmark datasets. The gains are particularly pronounced on Camelyon17-WILDS, where DINO and iBOT improve Top-1 accuracy by 6-10%. In contrast, improvements on ULiege and NCT-CRC-HE-100K are more limited. BYOL follows a similar pattern, except for its fine-tuned variant on Camelyon17-WILDS, which performs substantially worse. DINOv3 improves on ULiege after in-domain training but loses performance on the two benchmark datasets, especially on Camelyon17-WILDS. Overall, training from scratch and fine-tuning produce similar results, with two notable exceptions: fine-tuning is markedly worse than training from scratch for BYOL on Camelyon17-WILDS, whereas the opposite is observed for DINOv3.

Foundation models. Most domain-specific foundation models exhibit broadly consistent behavior. However, UNI (v1 and v2), Virchow v2, and H-optimus (0 and 1) achieve accuracies above 90% on Camelyon17-WILDS, whereas the remaining models generally fall within the 75-80% range. This difference may be explained by the composition of their pretraining datasets. While all models are trained on histopathological images, UNI, H-optimus, and Virchow v2 were developed using particularly diverse datasets spanning multiple tissues, organs, and pathologies, which likely contributes to their strong performance on both Camelyon17-WILDS and NCT-CRC-HE-100K. In contrast, models trained on less diverse data show larger performance variations across benchmarks. Retrieval performance generally decreases on the ULiege dataset, which covers multiple organs and pathologies and is therefore inherently more challenging than the more specialized Camelyon17-WILDS and NCT-CRC-HE-100K datasets. Among all models, cdpath21-gan clearly underperforms. Although it is the only model trained exclusively with a generative objective, its weaker results are more likely attributable to its smaller backbone (ResNet18) and more limited training dataset.

Evolution of the models. Our comparison includes subsequent iterations of several models and frameworks. Dinov3 deteriorates performance with respect to Dino, except on Camelyon17-WILDS. The mixed performances of Dinov3 on whole-slide images have been reported elsewhere [81]. Regarding the in-domain foundation models, no substantial performance differences can be observed between version 1 and version 2 of the models. While UNIv2 and Phikonv2 consistently outperform their respective v1 counterparts on the ULiege dataset, the opposite trend is observed on the NCT-CRC-HE-100K and Camelyon17-WILDS datasets. H-optimus-1 slightly outperforms H-optimus-0 across all metrics and datasets; however, the margin remains limited. Similarly, Virchow v2 achieves performance levels comparable to other models and does not demonstrate a clear advantage over earlier versions. It is also important to consider that newer versions of most models generally involve increased architectural complexity and higher-dimensional feature representations. While these modifications may yield modest performance improvements, they also introduce a non-negligible computational cost for retrieval (see Section 5.1). Consequently, the trade-off between accuracy gains and efficiency should be carefully evaluated when selecting a model.

Global comparison. The performance gap between out-of-domain and in-domain SSL models is relatively small, highlighting the strong generalization ability of large pretrained models. Nevertheless, the best-performing model belongs to the in-domain category, and in-domain SSL models maintain a slight overall advantage, particularly when trained

on large and diverse datasets. They should therefore be preferred when maximizing retrieval performance is the primary objective. General-purpose SSL models remain competitive, however, and represent a practical alternative when in-domain foundation models are unavailable, for example due to licensing constraints or limited computational resources.

Model architecture and training strategy also have a substantial impact on performance, with approaches such as DINO and iBOT consistently yielding strong results. In contrast, supervised models are advantageous mainly in in-distribution settings, where training and evaluation data share similar characteristics. They should therefore be considered only when labeled training data are available and retrieval is expected to be performed on data drawn from a comparable distribution.

5. Discussion

Although retrieval accuracy metrics such as Top-k and Maj-k are central for evaluating CBIR systems, successful deployment in digital pathology also depends on practical considerations. Retrieval systems must not only return diagnostically relevant cases but also operate with low latency and be distributed under licenses compatible with clinical or research use. In addition, retrieval performance should be interpreted in the context of how pathologists interact with such systems during routine diagnostic workflows.

While our study does not include a formal reader's study, the large-scale evaluation presented here provides valuable insights into the models most likely to be effective in practice. The following sections discuss three complementary aspects of real-world deployment: computational efficiency, licensing constraints, and the relationship between retrieval metrics and practical pathology workflows.

5.1. Computational Efficiency and Scalability

Retrieval latency is a key practical consideration for interactive systems. Table 4 reports processing times and embedding sizes for all evaluated models on the ULiege dataset, using 96,066 queries and an index of 106,281 images. Measurements were obtained under identical conditions on an NVIDIA RTX 4090 GPU and an Intel i9-13900KF CPU.

We report both total query time (feature extraction and search) and search time alone. Across all models, total latency remains below 30 ms per query, indicating that large-scale interactive retrieval is feasible. However, notable differences exist between approaches. Second-generation foundation models, such as UNI v2 and Phikon v2, are substantially slower than their predecessors, likely due to their larger architectures. In contrast, database search contributes only marginally to overall latency. Although retrieval time increases with embedding dimensionality (0.79 ms for 128 dimensions versus 1.9 ms for 2560 dimensions), feature extraction remains the main computational bottleneck.

Interestingly, the supervised models that performed best in our benchmarks are among the most computationally efficient approaches. Future studies involving pathology experts could help determine whether the performance gains offered by large-scale foundation models justify their increased computational demands in routine clinical workflows.

Model	Embedding vector size	Total Time/query (ms)	Search time/query (ms)
Supervised			
ResNet	128	5.66	0.79
DeiT	128	7.15	0.82
SSL Out-of-domain			
Dino	384	6.36	0.93

BYOL	256	5.80	0.85
iBOT	384	6.17	0.92
Dinov3	1024	9.6	1.19
SSL In-domain			
RetCCL	2048	6.58	1.69
CTransPath	768	6.92	1.07
cdpath21-gan	512	3.30	0.91
UNI	1024	11.60	1.20
UNI v2	1536	19.44	1.47
Phikon	768	7.00	1.10
Phikon v2	1024	13.50	1.21
H-optimus-0	1536	27.23	1.47
H-optimus-1	1536	27.14	1.47
Virchow	2560	21.04	1.90

Table 4. Total Time (ms) taken to process and retrieve Top-10 results for 1 query and search time (comparison of the feature vector of the query to the database) for each discussed model on the ULiege dataset.

5.2. Licensing and Reusability Constraints

Model and code licensing are also critical in determining whether an approach can realistically be integrated into real-world workflows. As shown in Table 1 (last column), the evaluated models are distributed under heterogeneous licensing conditions. Notably, some of the best-performing foundation models, including UNI v2 and H-optimus-1, are released under restrictive non-commercial licenses, with model weights available only upon request. Such constraints may limit large-scale deployment and integration into pathology platforms. In contrast, several other models provide openly accessible weights with more permissive licensing terms. In this regard, H-optimus-0 appears particularly interesting, as it achieves performance close to H-optimus-1 while remaining significantly easier to access and deploy.

5.3. Retrieval Metrics as Proxies for Clinical Utility

The practical usefulness of retrieval systems depends on how effectively they support human diagnostic reasoning. In typical workflows, pathologists may use CBIR systems to compare ambiguous visual patterns against previously observed or annotated cases. In this context, the relevance of the highest-ranked retrieved examples is particularly important, as users are unlikely to inspect large result sets during routine analysis. From this perspective, we believe ranking-based retrieval metrics used in this paper (Top-k and Maj-k accuracies) can be interpreted as proxies for the likelihood that diagnostically relevant visual evidence is presented early to the user as illustrated in the following example.

Illustrative use case in thyroid cytology. To demonstrate this principle, we applied our retrieval framework to annotated whole-slide images from routine thyroid cytology that were scanned (Nanozoomer HT 2.0, 40X, 0.23µm/pixel) and annotated by pathology experts. Figure 6 shows a query patch containing a papillary cell with an intranuclear inclusion, a key feature of papillary thyroid carcinoma [74]. The retrieval database consisted of 1,769 annotated cells, including both papillary cells with intranuclear inclusions and normal follicular cells. Given the query image, the system returned the ten most similar cells (Figure 7). Among these, eight corresponded to papillary cells with intranuclear inclusions. This high concentration of diagnostically consistent patterns among the top-ranked results provides robust evidence that the query cell belongs to the same category. As intranuclear inclusions are an important diagnostic

criterion for papillary thyroid cancer [74], this example illustrates how CBIR can support pathological assessment by quickly surfacing relevant visual evidence.

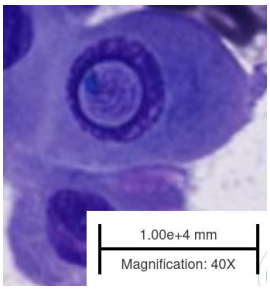

Figure 6: Patch containing a papillary cell with inclusion in thyroid cytology.

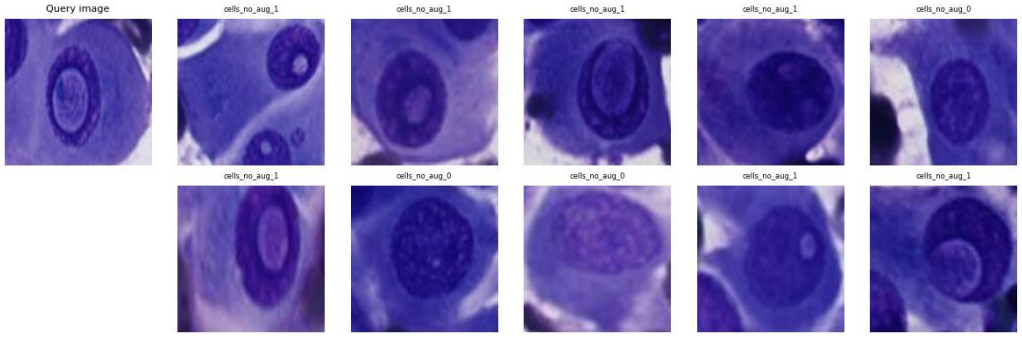

Figure 7: Top10 results of the framework using H-optimus-0 as backbone when using as query the patch of a papillary cell with an inclusion. Among the ten retrieved images, eight are representing cells previously annotated as papillary cells with intranuclear inclusion.

6. Conclusions

We compared 24 variants derived from 12 state-of-the-art feature extractors for patch-level image retrieval in digital pathology. Our objective was to evaluate the impact of training paradigm (supervised vs. self-supervised) and training data characteristics (in-domain vs. out-of-domain) on retrieval performance and generalization. Supervised models achieved the best results when evaluated on data drawn from the same distribution as their training set. However, their advantage over self-supervised models remained limited, while self-supervised approaches demonstrated superior generalization to out-of-distribution data. Among them, out-of-domain foundation models showed strong transferability, but in-domain models consistently achieved the best overall performance and should therefore be preferred in most settings. Our results also highlight the importance of training data, as models trained on larger and more diverse histopathology datasets were generally more robust and generalized better across datasets.

Quantitative retrieval metrics alone cannot fully capture clinical usefulness or diagnostic impact. Assessing factors such as diagnostic confidence, interpretability, and integration into routine practice would require dedicated reader studies involving pathologists. However, performing such evaluations across 24 models and multiple datasets would be impractical without first establishing a large-scale benchmark. In this context, we believe that the empirical study presented here, together with the accompanying open-source library, provides a valuable foundation for informed model selection and future targeted evaluations. Such studies may be further facilitated by the user-friendly retrieval interface developed for the Cytomine open-source platform.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

Statement: *During the preparation of this work the author(s) used Microsoft Copilot in order to improve the readability of the writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.*

7. References

[1] Wang X, Xiang J, Zhang J, Yang S, Yang Z, Wang MH, Zhang J, Yang W, Huang J, Han X. SCL-WC: cross-slide contrastive learning for weakly-supervised whole-slide image classification. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. Advances in Neural Information Processing Systems. Vol 35. Curran Associates, Inc.; 2022:18009-18021. https://doi.org/10.52202/068431-1309

[2] Qu L, Luo X, Wang M, Song Z. Bi-directional weakly supervised knowledge distillation for whole slide image classification. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. Advances in Neural Information Processing Systems. Vol 35. Curran Associates, Inc.; 2022:15368-15381. https://doi.org/10.52202/068431-1118

[3] Greenwald NF, Miller G, Moen E, Kong A, Kagel A, Dougherty T, Fullaway CC, McIntosh BJ, Leow KX, Schwartz MS, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nat Biotechnol. 2022;40(4):555-565. https://doi.org/10.1038/s41587-021-01094-0

[4] Xie R, Pang K, Chung S, Perciani C, MacParland S, Wang B, Bader G. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In: Oh A, Neumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. Advances in Neural Information Processing Systems. Vol 36. Curran Associates, Inc.; 2023:70626-70637. https://doi.org/10.52202/075280-3095

[5] Agarwal S, Eltigani Osman Abaker M, Daescu O. Survival prediction based on histopathology imaging and clinical data: a novel whole-slide CNN approach. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer; 2021:762-771. https://doi.org/10.1007/978-3-030-87240-3_73

[6] Chen RJ, Krishnan RG. Self-supervised vision transformers learn visual concepts in histopathology. LMRL 2021. Published 2022. https://arxiv.org/abs/2203.00585.

[7] Tu C, Zhang Y, Ning Z. Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. Advances in Neural Information Processing Systems. Vol 35. Curran Associates, Inc.; 2022:29484-29497. https://doi.org/10.52202/068431-2138

[8] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021:9630-9640. https://doi.org/10.1109/ICCV48922.2021.00951

[9] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020:9729-9738.https://doi.org/10.1109/CVPR42600.2020.00975

[10] Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, et al. Bootstrap your own latent: a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. Vol 33; 2020:21271-21284. https://doi.org/10.48550/arXiv.2006.07733

[11] Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: a survey. IEEE Trans Pattern Anal Mach Intell. 2021;43(11):4037-4058. https://doi.org/10.1109/TPAMI.2020.2992393

[12] Zhang T, Wei D, Zhu M, Gu S, Zheng Y. Self-supervised learning for medical image data with anatomy-oriented imaging planes. Med Image Anal. 2024;94:103151. https://doi.org/10.1016/j.media.2024.103151

[13] Wang X, Du Y, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. Med Image Anal. 2023;83:102645. https://doi.org/10.1016/j.media.2022.102645

[14] Kang M, Song H, Park S, Yoo D, Pereira S. Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023:3344-3354. https://doi.org/10.1109/CVPR52729.2023.00326

[15] 15. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Comput Sci. 2022;8:e1045. https://doi.org/10.7717/peerj-cs.1045

[16] Bulten W, Kartasalo K, Chen PHC, Ström P, Pinckaers H, Nagpal K, Cai Y, Steiner DF, van Boven H, Vink R, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med. 2022;28(1):154-163. https://doi.org/10.1038/s41591-021-01620-2

[17] Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-1120. https://doi.org/10.1038/ng.2764

[18] Bandi P, Geessink O, Manson Q, van Dijk M, Balkenhol M, Hermsen M, Bejnordi BE, Lee B, Paeng K, Zhong A, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans Med Imaging. 2019;38(2):550-560. https://doi.org/10.1109/TMI.2018.2867350

[19] 19. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. Vol 139. Proceedings of Machine Learning Research; 2021:8748-8763. https://arxiv.org/abs/2103.00020

[20] Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, Kong T. iBOT: image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (ICLR); 2022. https://arxiv.org/abs/2111.07832

[21] Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, Chen B, Zhang A, Shao D, Shaban M, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850-862. https://doi.org/10.1038/s41591-024-02857-3

[22] Filiot A, Ghermi R, Olivier A, Jacob P, Fidon L, Mac Kain A, Saillard C, Schiratti JB. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv. Published July 26, 2023. https://doi.org/10.1101/2023.07.21.23292757

[23] Saillard C, Jenatton R, Llinares-López F, Mariet Z, Cahané D, Durand E, Vert JP. H-optimus-0. Published 2024. Accessed September 30, 2026. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0

[24] Boyd J, Liashuha M, Deutsch E, Paragios N, Christodoulidis S, Vakalopoulou M. Self-supervised representation learning using visual field expansion on digital pathology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops; 2021:639-647. https://arxiv.org/abs/2109.03299

[25] Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal. 2022;81:102559. https://doi.org/10.1016/j.media.2022.102559

[26] Zimmermann E, Vorontsov E, Viret J, Casson A, Zelechowski M, Shaikovski G, Tenenholtz N, Hall J, Fuchs T, Fusi N, Liu S, Severson K. Virchow2: scaling self-supervised mixed magnification models in pathology. arXiv. Published 2024. https://doi.org/10.48550/arXiv.2408.00738

[27] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009:248-255. https://doi.org/10.1109/CVPR.2009.5206848

[28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. Published 2015. https://doi.org/10.48550/arXiv.1512.03385

[29] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: transformers for image recognition at scale. arXiv. Published 2021. https://doi.org/10.48550/arXiv.2010.11929

[30] Elharrouss O, Akbari Y, Almadeed N, Al-Maadeed S. Backbones-review: feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. Comput Sci Rev. 2024;53:100645. https://doi.org/10.1016/j.cosrev.2024.100645

[31] Wu H, Wang M, Zhou W, Li H. Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021:11416-11425. https://doi.org/10.1109/ICCV48922.2021.01122

[32] Zhong A, Li X, Wu D, Ren H, Kim K, Kim Y, Buch V, Neumark N, Bizzo B, Tak WY, et al. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in COVID-19. Med Image Anal. 2021;70:101993. https://doi.org/10.1016/j.media.2021.101993

[33] Li B, Lima D. Facial expression recognition via ResNet-50. Int J Cogn Comput Eng. 2021;2:57-64. https://doi.org/10.1016/j.ijcce.2021.02.002

[34] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Guyon I, van Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. Advances in Neural Information Processing Systems. Vol 30. Curran Associates, Inc.; 2017:5998-6008. https://doi.org/10.48550/arXiv.1706.03762

[35] Zhang J, Zhang X, Ma K, Gupta R, Saltz J, Vakalopoulou M, Samaras D. Gigapixel whole-slide images classification using locally supervised learning. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science. Vol 13432. Springer Nature Switzerland; 2022:192-201. https://doi.org/10.1007/978-3-031-16434-7_19

[36] Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, Zhang Y. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, eds. Advances in Neural Information Processing Systems. Vol 34. Curran Associates, Inc.; 2021:2136-2147. https://doi.org/10.48550/arXiv.2106.00908

[37] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Daumé H III, Singh A, eds. Proceedings of the 37th International Conference on Machine Learning. Vol 119. Proceedings of Machine Learning Research; 2020:1597-1607. https://doi.org/10.48550/arXiv.2002.05709

[38] Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems. Vol 33; 2020:9912-9924. https://doi.org/10.48550/arXiv.2006.09882

[39] Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y. Learning fine-grained image similarity with deep ranking. CoRR. 2014;abs/1404.4661. https://doi.org/10.48550/arXiv.1404.4661

[40] Ng T, Balntas V, Tian Y, Mikolajczyk K. SOLAR: second-order loss and attention for image retrieval. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV. Lecture Notes in Computer Science. Vol 12370. Springer; 2020:253-270. https://doi.org/10.1007/978-3-030-58595-2_16

[41] Subramanian M, Lingamuthu V, Venkatesan C, Perumal S, et al. Content-based image retrieval using colour, gray, advanced texture, shape features, and random forest classifier with optimized particle swarm optimization. Int J Biomed Imaging. 2022;2022:3211793. https://doi.org/10.1155/2022/3211793

[42] Yang M, He D, Fan M, Shi B, Xue X, Li F, Ding E, Huang J. DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021:11772-11781. https://doi.org/10.1109/ICCV48922.2021.01156

[43] Deng Z, Zhong Y, Guo S, Huang W. InsCLR: improving instance retrieval with self-supervision. Proc AAAI Conf Artif Intell. 2022;36(1):516-524. https://doi.org/10.1609/aaai.v36i1.19930

[44] Hegde N, Hipp JD, Liu Y, Emmert-Buck MR, Reif E, Smilkov D, Terry M, Cai CJ, Amin MB, Mermel CH, Nelson PQ, Peng LH, Corrado GS, Stumpe MC. Similar image search for histopathology: SMILY. NPJ Digit Med. 2019;2:56. https://doi.org/10.1038/s41746-019-0131-z

[45] Kalra S, Tizhoosh HR, Choi C, Shah S, Diamandis P, Campbell CJV, Pantanowitz L. Yottixel: an image search engine for large archives of histopathology whole slide images. Med Image Anal. 2020;65:101757. https://doi.org/10.1016/j.media.2020.101757

[46] Chen C, Lu MY, Williamson DFK, Chen TY, Schaumberg AJ, Mahmood F. Fast and scalable search of whole-slide images via self-supervised deep learning. Nat Biomed Eng. 2022;6(12):1420-1434. https://doi.org/10.1038/s41551-022-00929-8

[47] Pantanowitz L, Michelow P, Hazelhurst S, Kalra S, Choi C, Shah S, Babaie M, Tizhoosh HR. A digital pathology solution to resolve the tissue floater conundrum. Arch Pathol Lab Med. 2021;145(3):359-364. https://doi.org/10.5858/arpa.2020-0034-OA

[48] Herdiantoputri RR, Komura D, Ochi M, Fukawa Y, Oba K, Tsuchiya M, Kikuchi Y, Matsuyama Y, Ushiku T, Ikeda T, Ishikawa S. Preclinical evaluation of an interactive image search system of oral pathology. J Dent Res. 2025;104(11):1192-1201. https://doi.org/10.1177/00220345251329042

[49] Kurmi Y, Chaurasia V. Content-based image retrieval algorithm for nuclei segmentation in histopathology images. Multimed Tools Appl. 2021;80(2):3017-3037. https://doi.org/10.1007/s11042-020-09797-3

[50] Wölflein G, Ferber D, Meneghetti AR, El Nahhas OSM, Truhn D, Carrero ZI, Harrison DJ, Arandjelović O, Kather JN. Benchmarking pathology feature extractors for whole slide image classification. arXiv. Published 2024. https://doi.org/10.48550/arXiv.2311.11772

[51] Shang HH, Nasr MS, Veerla JP, Saurav JR, Hajighasemi A, Malidarreh P, Huber M, Moleta C, Makker J, Luber JM. Histopathology slide indexing and search: are we there yet? NEJM AI. 2024;1(5):AIcs2300019. https://doi.org/10.1056/AIcs2300019

[52] Li S, Zhao Y, Zhang J, Yu T, Zhang J, Gao Y. High-order correlation-guided slide-level histology retrieval with self-supervised hashing. IEEE Trans Pattern Anal Mach Intell. 2023;45(9):11008-11023. https://doi.org/10.1109/TPAMI.2023.3269810

[53] Neidlinger P, El Nahhas OSM, Muti HS, Lenz T, Hoffmeister M, Brenner H, van Treeck M, Langer R, Dislich B, Behrens HM, Röcken C, Foersch S, Truhn D, Marra A, Saldanha OL, Kather JN. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv. Published 2024. https://doi.org/10.48550/arXiv.2408.15823

[54] Kim YJ, Jang H, Lee K, Park S, Min SG, Hong C, Park JH, Lee K, Kim J, Hong W, Jung H, Liu Y, Rajkumar H, Khened M, Krishnamurthi G, Yang S, Wang X, Han CH, Kwak JT, Ma J, Tang Z, Marami B, Zeineh J, Zhao Z, Heng PA, Schmitz R, Madesta F, Rösch T, Werner R, Tian J, Puybareau E, Bovio M, Zhang X, Zhu Y, Chun SY, Jeong WK, Park P, Choi J. PAIP 2019: liver cancer segmentation challenge. Med Image Anal. 2021;67:101854. https://doi.org/10.1016/j.media.2020.101854

[55] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Meila M, Zhang T, eds. Proceedings of the 38th International Conference on Machine Learning. Vol 139. Proceedings of Machine Learning Research. PMLR; 2021:10347-10357. https://proceedings.mlr.press/v139/touvron21a.html

[56] Mohan DD, Jawade B, Setlur S, Govindaraj V. Deep metric learning for computer vision: a brief overview. arXiv. Published 2023. https://doi.org/10.48550/arXiv.2312.10046

[57] Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Liu S, Severson K, Zimmermann E, Hall J, Tenenholtz N, Fusi N, Mathieu P, van Eck A, Lee D, Viret J, Robert E, Wang YK, Kunz JD, Lee MCH, Bernhard J, Godrich RA, Oakley G, Millar E, Hanna M, Retamero J, Moye WA, Yousfi R, Kanan C, Klimstra D, Rothrock B, Fuchs TJ. Virchow: a million-slide digital pathology foundation model. arXiv. Published 2024. https://doi.org/10.48550/arXiv.2309.07778

[58] Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, Lomeli M, Hosseini L, Jégou H. The Faiss library. arXiv. Published 2024. https://doi.org/10.48550/arXiv.2401.08281

[59] Mormont R, Geurts P, Marée R. Comparison of deep transfer learning strategies for digital pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018:2262-2271. https://ieeexplore.ieee.org/document/8575475

[60] Mormont R, Geurts P, Marée R. Multi-task pre-training of deep neural networks for digital pathology. IEEE J Biomed Health Inform. 2021;25(2):412-421. https://doi.org/10.1109/JBHI.2020.2992878

[61] Ehteshami Bejnordi B, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, Hermsen M, Manson QF, Balkenhol M, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199-2210. https://doi.org/10.1001/jama.2017.14585

[62] Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Training echo state networks for rotation-invariant bone marrow cell classification. Neural Comput Appl. 2017;28(6):1277-1292. https://doi.org/10.1007/s00521-016-2609-9

[63] Marée R, Dallongeville S, Olivo-Marin JC, Meas-Yedid V. An approach for detection of glomeruli in multisite digital pathology. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016:1033-1036. https://doi.org/10.1109/ISBI.2016.7493442

[64] Veta M, Heng YJ, Stathonikos N, Ehteshami Bejnordi BE, Beca F, Wollmann T, Rohr K, Shah MA, Wang D, Rousson M, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med Image Anal. 2019;54:111-121. https://doi.org/10.1016/j.media.2019.02.012

[65] Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M, Fernandez G, Zeineh J, Kohl M, Walz C, Ludwig F, Braunewell S, Baust M, Vu QD, To MNN, Kim E, Kwak JT, Galal S, Sanchez-Freire V, Brancati N, Frucci M, Riccio D, Wang Y, Sun L, Ma K, Fang J, Kone I, Boulmane L, Campilho A, Eloy

C, Polónia A, Aguiar P. BACH: Grand challenge on breast cancer histology images. Med Image Anal. 2019;56:122-139. https://doi.org/10.1016/j.media.2019.05.010

[66] Sirinukunwattana K, Raza SEA, Tsang YW, Snead DRJ, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging. 2016;35(5):1196-1206. https://doi.org/10.1109/TMI.2016.2525803

[67] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform. 2016;7(1):29. https://doi.org/10.4103/2153-3539.186902

[68] Kather JN, Halama N, Marx A. 100,000 histological images of human colorectal cancer and healthy tissue [dataset]. Zenodo. Published April 7, 2018. https://doi.org/10.5281/zenodo.1214456

[69] Roux L, Racoceanu D, Capron F, Calvo J, Attieh E, Le Naour G, Gloaguen A. Mitos & atypia. Singapore: Image Pervasive Access Lab (IPAL), Agency for Science, Technology and Research (A*STAR), Institute for Infocomm Research; Technical Report 1. Available at: http://ludo17.free.fr/mitos_atypia_2014/icpr2014_MitosAtypia_DataDescription.pdf. Accessed September 30, 2026.

[70] Linder N, Konsti J, Turkki R, Rahtu E, Lundin M, Nordling S, Haglund C, Ahonen T, Pietikäinen M, Lundin J. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagn Pathol. 2012;7:22. https://doi.org/10.1186/1746-1596-7-22

[71] Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, et al. WILDS: a benchmark of in-the-wild distribution shifts. In: Meila M, Zhang T, eds. Proceedings of the 38th International Conference on Machine Learning. Vol 139. Proceedings of Machine Learning Research. PMLR; 2021:5637-5664. https://proceedings.mlr.press/v139/koh21a.html

[72] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579-2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

[73] Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, Begon JM, Kainz P, Geurts P, Wehenkel L. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016;32(9):1395-1401. https://doi.org/10.1093/bioinformatics/btw013

[74] Harahap AS, Jung CK. Cytologic hallmarks and differential diagnosis of papillary thyroid carcinoma subtypes. J Pathol Transl Med. 2024;58(6):265-282. https://doi.org/10.4132/jptm.2024.10.11

[75] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv. Published 2018. https://doi.org/10.48550/arXiv.1807.03748

[76] Bachman P, Hjelm RD, Buchwalter W. Learning representations by maximizing mutual information across views. In: Advances in Neural Information Processing Systems. Vol 32. Curran Associates Inc; 2019. https://doi.org/10.48550/arXiv.1906.00910

[77] Roth K, Milbich T, Sinha S, Gupta P, Ommer B, Cohen JP. Revisiting training strategies and generalization performance in deep metric learning. In: Daumé H III, Singh A, eds. Proceedings of the 37th International Conference on Machine Learning. Vol 119. Proceedings of Machine Learning Research. PMLR; 2020:8242-8252. https://proceedings.mlr.press/v119/roth20a.html

[78] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Dy J, Krause A, eds. Proceedings of the 35th International Conference on Machine Learning. Vol 80. Proceedings of Machine Learning Research. PMLR; 2018:2127-2136. https://proceedings.mlr.press/v80/ilse18a.html

[79] Kaya M, Bilge HS. Deep metric learning: a survey. Symmetry (Basel). 2019;11(9):1066. https://doi.org/10.3390/sym11091066

[80] Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C, et al. DINOv3. arXiv. Published August 13, 2025. https://doi.org/10.48550/arXiv.2508.10104

[81] Liu C, Chen Y, Shi H, Lu J, Jian B, Pan J, et al. Does DINOv3 set a new medical vision standard? Benchmarking 2D and 3D classification, segmentation, and registration. arXiv. Published September 8, 2025. https://doi.org/10.48550/arXiv.2509.06467