Harnessing Self-Supervised Features for Histopathological Image Retrieval and Classification Through Efficient Fine-Tuning

José Solenir L. Figuerêdo[1], **Luciano Araújo D. Filho**[1], **Rodrigo Tripodi Calumby**[1]

[1]Postgraduate Program in Computer Science
University of Feira de Santana (UEFS)
Feira de Santana – BA – Brazil

{solenir.figueredo,lucianoadfilho}@gmail.com, rtcalumby@uefs.br

Abstract. *This study evaluates machine learning models pre-trained using the self-supervised architectures iBOT and DINOv2 for disease/lesion classification and histopathological image retrieval. Experiments were conducted using five datasets and explored three fine-tuning strategies: Full, Low-Rank Adaptation (LoRa), and Linear Probe. The LoRa technique yielded significant effectiveness gains, improving model effectiveness by over 40% in some scenarios (LoRa vs Full). In image retrieval, the DINO-Hist model outperformed iBOT-Hist, demonstrating statistically significant superiority in MAP@10 and MAP@40. These findings underscore the adaptability of self-supervised architectures and the critical role of fine-tuning in enhancing model effectiveness.*

1. Introduction

Histopathology involves the microscopic analysis of biological cells and tissue structures to diagnose diseases such as cancer, liver, and kidney disorders [Abels et al. 2019, Chagas et al. 2020, L'Imperio et al. 2021, Cerqueira et al. 2021]. Despite its effectiveness, this manual evaluation is labor-intensive, subjective, and prone to errors. To enhance diagnostic accuracy and reduce pathologists' workload, computer-aided diagnosis systems have emerged as a promising solution. Artificial intelligence applications leverage collected images to train disease detection models and facilitate content-based image retrieval (CBIR) of similar cases.

In recent years, deep learning-based CBIR systems have been developed, relying on large labeled datasets for feature extraction. However, medical data labeling remains a significant challenge, particularly in pathologies requiring meticulous annotation, such as renal diseases, where identifying all glomeruli is time-consuming and error-prone [Yamaguchi et al. 2021].

To address the challenges of data labeling in digital pathology, studies have explored transfer learning from ImageNet [Ozen et al. 2021, Majumdar et al. 2023] and, more recently, self-supervised learning (SSL), which extracts semantic features from unlabeled data [Chen et al. 2019]. Techniques such as SwAV, iBOT, and EsVIT [Balestriero et al. 2023] enable pre-training on large datasets, followed by fine-tuning on tasks with limited annotations. However, models trained on natural images may not fully capture the complexity of histological structures [Kataria et al. 2023]. To overcome this, Filiot et al. (2023) developed iBOT-based models pre-trained on histopathological images, achieving superior results across various tasks. Despite their potential, these models have not been evaluated in CBIR tasks or specific histological domains like nephropathology, where detailed glomerular analysis is crucial. Their application could enhance effectiveness in tasks with limited labeled data, making them valuable for specialized medical image analysis.

Despite significant improvement of SSL techniques in the last years, training models from random weights can be very restrictive since its pre-training demands significant amounts of computational resources and may not always be feasible in low data regimes. To overcome this limitation, foundation models have emerged as an alternative to SSL from scratch. Recently, one of these models, called DINOv2 [Oquab et al. 2023], attracted the interest of the scientific community. DINOv2 is an open-source foundation model pre-trained through self-supervised learning and has achieved state-of-the-art results in several tasks. DINOv2 outperformed several alternative methods, including the iBOT in a wide range of benchmarks [Oquab et al. 2023]. However, despite achieving state-of-the-art results, questions still remain regarding the adaptability of DINOv2 to histopathological images, especially renal images. Furthermore, its application in CBIR tasks still requires further analysis.

Thus, considering the aforementioned challenges, in this study we developed and experimentally evaluated two self-supervised models applied in the context of histopathological images. One of these models was pre-trained with domain images, while the other was pre-trained with natural images. We conducted a comparative analysis of three fine-tuning techniques, evaluating the impact of each technique on the effectiveness of the results, in the context of CBIR and also classification.

2. Related Works

In a CBIR system, images that share visual attributes, such as color, texture, and shape, with a query image are indexed and retrieved from large databases. In this context, in recent years, several works have been developed for this purpose, often using more advanced Artificial Intelligence approaches, such as deep learning and self-supervised learning. Among the relevant recent works, the following stand out: [Yang et al. 2020], [Zheng et al. 2022], [Mohammad Alizadeh et al. 2023], [Wickstrøm et al. 2023] and [Filiot et al. 2023].

Yang et al. (2020) designed an image search engine for digital pathology, but focused on representing WSIs in a compact way. The authors developed an indexing algorithm to represent WSIs as a mosaic of patches that are then converted to barcodes, called "Bunch of Barcodes" (BoB). The authors evaluated the approach using a *ground-truth*, as well as a subjective evaluation, with expert and non-expert users of the domain. The results found were promising, indicating that the engine can accurately retrieve organs and malignancies, and its semantic ordering shows effective agreement with the subjective evaluation of human observers and the search engine. In Zheng et al. (2022), the authors also proposed an approach for representing WSI. A graph-based framework was designed to enable region-of-interest searching, complemented by a feature encoding process for binarization. The obtained results demonstrate the superiority of the proposed method over comparable approaches.

In Mohammad Alizadeh et al. (2023), the authors proposed a novel hashing-based Siamese convolutional neural network method for histopathological image retrieval based on a pairwise structure. The authors also use a binary encoding of the features. For this, two deep hashing models with shared weights and structures are used. A new cost function is also designed to improve the training and image retrieval process. The method was evaluated on two public datasets, and according to the experimental results, the developed model outperformed other hashing-based methods. In a distinct study, Wickstrøm et al.(2023) proposed a self-supervised framework for liver CT image retrieval, diverging from traditional methodologies. Additionally, acknowledging the critical importance of explainability in medical applications, the authors conducted the first explainability analysis of representation learning within the context of content-based image retrieval (CBIR) for liver computed tomography images. The results indicate that the self-supervised approach allowed to extract clinically relevant features, and the practical usability of the proposed framework was also validated by an expert.

Filiot et al. (2023) explores the application of Masked Image Modeling (MIM) to histological images using the ViT-based iBOT framework. By training a model to predict hidden or masked parts of an image, MIM allows a computer vision model to extract meaningful representations from images. In the work of [Filiot et al. 2023], the pre-trained model was evaluated on several downstream tasks, all of them related to cancer. Although the authors did not use the pre-trained model in the context of CBIR systems, they made great contributions. One of these contributions refers to the availability of the pre-trained models to serve as a base model for future studies.

3. Experimental Pipeline

The experimental process conducted in this study is illustrated in Figure 1. It consists of 6 steps: (1) Data collection, (2) Dataset partitioning, (3) Preprocessing, (4) Adaptation of pre-trained models (fine-tuning), (5) Extraction of feature vectors and, finally, (6) Model evaluation in classification and retrieval tasks.

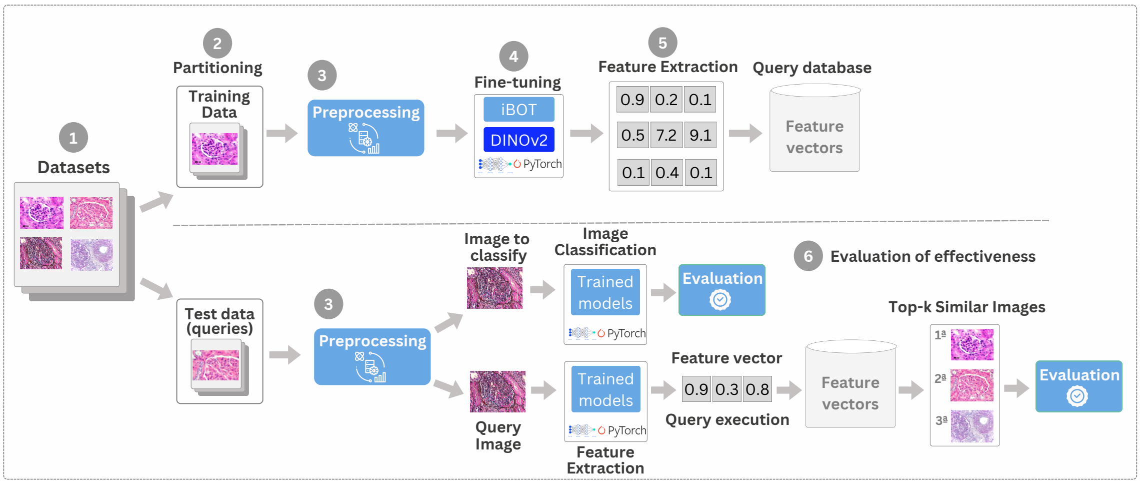


Figure 1. Experimental Benchmark Workflow.

3.1. Datasets and Partitioning

Five datasets were used in the experimental process, covering different types of histological tissues and different pathologies or associated lesions. In this study, the datasets were named as: AIDPATH, PathoSpotter-HE, PathoSpotter-PAS, PathoSpotter-MultiContrast and RCC. Each of these databases is detailed below.

- **AIDPATH:** This database is part of the WSI datasets generated in the European project AIDPATH[1]. The database contains 2340 images with a single glomerulus, which are used for the development of studies related to the identification of Glomerulosclerosis [Bueno et al. 2020]. Of the 2340 images, 1170 show normal glomeruli, while the remaining 1170 show sclerotic glomeruli.
- **PathoSpotter-HE:** It is part of the *PathoSpotter*[2] project, linked to the Oswaldo Cruz Foundation (Fiocruz) of Bahia. This particular *dataset* has 4947 images, divided into six classes of interest: Normal (869), Hypercellularity (1237), Primary membranous (712), Secondary membranous (1354), Sclerosis with membranous (264) and Sclerosis without membranous (511). To improve the visualization of specific structures, the Hematoxylin and Eosin (H&E) stain was used.
- **PathoSpotter-PAS:** This *dataset* is also part of the PathoSpotter project, but unlike PathoSpotter-HE, it uses Periodic Acid-Schiff (PAS) as a stain. This set has 2390 images, and is also divided into six classes of interest: Normal (293), Hypercellularity (637), Primary membranous (367), Secondary membranous (609), Sclerosis with membranous (156) and Sclerosis without membranous (328).
- **PathoSpotter-MultiStain:** This dataset is also part of the PathoSpotter project, but uses different dyes. In addition to having images that used H&E and PAS, it also has images that used PAMS (Picroaniline Methenamine Silver), PS (Picro-Sirius Red), AZAN (Alcian Blue and Orange) and PICRO (Picro-Chromic Weigert). The database is composed of 10184 images, also covering six classes: Normal (1570), Hypercellularity (2043), Primary membranous (1608), Secondary membranous (3170), Sclerosis with membranous (513) and Sclerosis without membranous (1280).
- **RCC:** This dataset contains images of renal cell carcinoma, obtained from different patients and severity grades, as part of a clinical study in the Department of Pathology, Kasturba Medical College (KMC), Manipal, India. The dataset includes non-cancerous (Grade 0) and cancerous (Grade 1 to Grade 4) images of the carcinoma. The samples were collected by surgical (open) biopsy of renal tissue and stained with H&E. The dataset consists of 4077 images, subdivided into 5 classes or grades of cancer: Grade 0 (836), Grade 1 (839), Grade 2 (769), Grade 3 (861) and Grade 4 (772).

Another step of this study consisted of partitioning the datasets before performing preprocessing. The data were partitioned into training, validation, and testing, following the proportion of 70:20:10, respectively. The only exception was the RCC dataset. For RCC, which was already partitioned into training, validation, and testing, it was decided to maintain the original division, even though it did not follow the proportion of 70:20:10. Table 1 presents a detailed description of the number of samples in the sets after performing the partitioning.

[1] `https://aidpath.eu/`

[2] `https://pathospotter.bahia.fiocruz.br/`

Table 1. Datasets Statistics

Dataset	Number of Classes	Stain	Train	Validation	Test	Total
AIDPATH	2	HE	1638	468	234	2340
PathoSpotter-HE	6	HE	3462	990	495	4947
PathoSpotter-PAS	6	PAS	1673	478	239	2390
PathoSpotter-MultiStain	6	HE, PAS, PAMS, PS, AZAN, PICRO	7128	2038	1018	10184
RCC	5	HE	3432	503	142	4077

3.2. Preprocessing

After the partitioning step, a preprocessing pipeline was conducted. The main goal of this step was to improve the training quality by introducing variability, as well as ensuring that the data is in the correct format. For the training set, the following pipeline was followed:

- **Resizing:** All images were scaled to the dimensions of 224 x 224 *pixels*, as required by the architecture of the network used, which was designed to operate with inputs of this specific size.
- **Random horizontal flip:** A random horizontal flip is applied to the images with a default probability of 50%, aiming to increase the robustness of the model with respect to different image orientations.
- **Conversion to RGB:** Converts the image to RGB color mode if it is not already in that format, ensuring that the image is processed with three color channels.
- **Normalization of pixel values:** A normalization is applied to the values of the *pixels* of the image using means and standard deviations specific to each color channel (R, G, B).

The same procedures were applied to the validation and test sets, except for the random horizontal flip.

3.3. Fine-tuning and Feature Extraction

One of the goals of this study is to evaluate the application of self-supervised pre-trained models in the context of histopathological image retrieval. To this end, one of the steps of this study consisted of the fine-tuning process of these models. Adaptation, or fine-tuning, refers to the additional adjustment of a model previously trained on a large database with the aim of adapting it to a new dataset or specific task. Evaluating different fine-tuning techniques is important for several reasons, such as optimized performance, generalization, computational efficiency, and domain adaptation. Considering the aforementioned characteristics, in this study three fine-tuning approaches were evaluated, namely:

- **Full:** In this approach, the weights of all layers are updated, which results in a higher consumption of computational resources. In some scenarios, this form of *fine-tuning* may be infeasible, depending on the available computational resources.
- **LoRA:** Low-Rank Adaptation or LoRA[3] is a model adaptation technique that provides an efficient way to fine-tune large-scale models for specific tasks. With

[3] https://huggingface.co/docs/peft/package_reference/lora

LoRA, the pre-trained model weights are frozen, and trainable rank decomposition matrices are injected into the transformer attention blocks, thereby reducing the number of trainable parameters. LoRa is particularly noteworthy among fine-tuning strategies due to its parameter efficiency and adaptability. Unlike full strategy, which adjusts all the parameters of the base model, LoRa focuses on injecting low-rank adaptations into specific layers of the architecture, typically in the attention projections

- **Linear Probe:** In this approach, a linear layer (classification head) is attached to the backbone model and only the classification layer weights are updated, whereas the remaining (backbone) weights remains frozen. Although this approach generally may not be regarded as fine-tuning per se, we will consider it as a lightweight form of fine-tuning, for practical purposes.

After fine-tuning, these feature extraction models, especially the one pre-trained in the histological domain, are expected to be able to capture both low-level features and semantic information. Considering the architectures employed, each feature vector presented 768 dimensions. These vectors were stored in a database for subsequent content-based image retrieval evaluation.

3.4. Experimental Setup

For experimental evaluation, two SSL pre-trained models were selected: the model developed in the work of [Filiot et al. 2023] and the base model DINOv2 [Baharoon et al. 2024]. Our study is focused exclusively on iBOT and DINOv2. The first model explores the application of MIM to histological images using the iBOT, ViT-based SSL pre-training method. One of the contributions of the study was the generation and provision of pre-trained models with histopathological images to be used as a base framework for future research in the area. On the other hand, DINOv2 was pre-trained with natural images, also using ViT. Although there are many base models trained through self-supervised learning with natural images, DINOv2 was selected due to the robustness of its representations, which has achieved competitive performance in multiple tasks, for different types of media (video and image).

The model generation process uses the training and validation dataset in the *fine-tuning* phase. The test set is used to evaluate the quality of the representations extracted from the models. It should be noted that all images are used at the *patch*[4] level. For each *dataset*, the three *fine-tuning* techniques described in section 3.3 are evaluated. The techniques called "Full" and "Linear Probe" do not require additional configuration, but it was necessary to specify some hyperparameters for the LoRa technique. Specifically, the following hyperparameters were configured: $r = 16$, $lora\_alpha = 16$, $target\_modules = [``query'', ``value'']$, $lora\_dropout = 0.1$, and $bias = ``none''$.

The fine-tuning experiments were conducted using 50 epochs, with the following batch variation: 16, 32, 64. In addition, three different learning rate values were used: 1×10^{-3}, 5×10^{-3} and 5×10^{-4}. The application of *weight decay* with a value of 0.01 was also evaluated compared to its absence. The other hyperparameters followed the default configuration established by TrainingArguments [5]. The experiments were performed for

[4] Refers to a smaller, rectangular or square subsection of the WSI.

[5] It is a class of the transformers package of the Hugging Face library, used to define and configure the hyperparameters for training and evaluating machine learning models.

each dataset, architecture, fine-tuning technique and hyperparameter variation. Thus, a total of 756 models were generated. For practical purposes, in this study we will present the most effective models only, i.e., the ones that presented the best F_1 Macro measure (from which a detailed description can be found in section 3.5).

In the CBIR evaluation process, training and validation sets were combined to form a single database for image retrieval assessment. On the other hand, the set of image queries was obtained from the remaining test set images, which corresponds to the set of images used to verify the quality of the extracted representations.

All experiments were performed using a single GPU NVIDIA Quadro P4000 of 20GB. The models adapted in this study have been made publicly available at the following electronic address: `https://drive.google.com/drive/folders/1zpaI5PTFtJCZOIxw0VMbdwjmyz6weoMm?usp=sharing`.

3.5. Evaluation

The classification models were evaluated using the following measures:

- **Accuracy:** Corresponds to the proportion of correct predictions made by the model in relation to the total of predictions.
- **Macro Precision:** *Precision* corresponds to the ratio of true positives in relation to the total of predicted positives, whereas Macro Precision corresponds to the (macro) average precision over all classes. In other words, each class precision is computed individually and then the macro average is obtained by taking the average precision over the total amount of classes.
- **Macro Recall:** The *recall* is the proportion of true positives in relation to the total amount of positives (true and false positives), whereas Macro Recall corresponds to the macro average of recall. This metric is capable of evaluating the coverage of the model.
- **Macro F_1:** This metric corresponds to the harmonic mean between *precision* and *recall*, providing a single metric that balances both measures, thereby the Macro F_1 corresponds to the macro average F_1.

In the context of content-based image retrieval, the *Mean Average Precision (MAP) at K* or simply $MAP@K$ was used for evaluation. It computes the mean of the Average Precisions (AP) across multiple queries, where AP quantifies the quality of the ranking of relevant results within a retrieved set. This measure allows to evaluate both the relevance of the ranked images and the effectiveness of the system in positioning the relevant images in the first positions. In this work, the relevant images are those that belong to the class of the query image.

The calculation of similarity between the images was performed using cosine similarity. Cosine similarity is a metric used to measure the degree of similarity between two vectors in a multidimensional space. Additionally, for strict comparison of efficacy results, the models developed were compared using the Wilcoxon test in order to assess statistical significance. It should be noted that several queries were executed, therefore, the value reported by $MAP@K$ corresponds to the average of these executions.

4. Results and discussion

The results are organized in two sections. A Section 4.1 discusses the effectiveness of the models in the classification task. On the other hand, Section 4.2 demonstrates the result for content-based image retrieval. From this point on, models trained with the base iBot architecture will be referred to as iBOT-Hist, while models trained with DINOv2 will be referred to as DINO-Hist.

4.1. Classification Effectiveness

Table 2 presents the effectiveness of the models evaluated in this study, taking into account each dataset and the fine-tuning variants. The numerically superior results are highlighted in bold for each dataset. In general, it is observed that the evaluated models achieved promising results in all evaluated measures. The most effective model for each dataset obtained a performance greater than or equal to 81%, reaching 100% effectiveness in some datasets (AIDPATH and Kather). These results demonstrate the effectiveness of the employed architectures and emphasize the importance and robustness of self-supervised learning. Ultimately, the knowledge acquired during the pre-training phase exerts a significant influence on the model's adjustment to other domains. In general, the DINO-Hist was superior to the iBOT-Hist in several scenarios. These results show that DINOv2 produces more robust and generalizable visual representations, effectively capturing the underlying semantic structure of images in a self-supervised manner.

Significant evidence from this study is associated with the fine-tuning techniques employed. In all scenarios evaluated, the models that used the LoRa technique were the ones that showed the highest efficiency. This result is of great relevance, since even without requiring many computational resources, LoRa provided significant improvements in efficiency. Additionally, the fact that the LoRa technique achieved the best efficiency in all datasets evaluated indicates its high efficiency in adapting models, providing high-quality results in several tasks or domains. The fact that it achieved the best results demonstrates its ability to generalize effectively in new data and domains, which is essential for the practical application of models in data from varied and unprecedented contexts, different from those used in the pre-training phase. This evidence becomes even more relevant when considering DINO-Hist, which was pre-trained with natural images. LoRa allowed it to be adapted to the domain of histopathological images, outperforming the Full fine-tuning and Linear Probe.

To facilitate understanding of the impact of LoRa, Table 3 presents the percentage gain of this technique compared to the others. The PathoSpotter-HE dataset was used as a reference. It can be seen that LoRa obtained gains above 9% for all evaluated measures, and this gain was even greater in DINO-Hist. For DINO-Hist, the LoRa technique obtained gains above 28% for all measures. The fact that the DINO-Hist base model was pre-trained with natural images may have influenced the observed results. Fine-tuning between completely different domains, such as from natural to histopathological images, does not always achieve the expected effectiveness. This challenge further highlights the importance of this technique, which has proven to be more effective than the other variants. In DINO-Hist, LoRa's superiority is even more expressive when compared to Full fine-tuning. This result is quite positive, since in addition to being more computationally expensive, the Full method had lower performance. The superior efficiency achieved by LoRa highlights its relevance as an efficient and effective approach for model adaptation. The intrinsic advantages of this technique qualify it as a valuable tool for model adaptation in various tasks and applications.

Table 2. Effectiveness of the models evaluated in this study

Dataset	Model	Fine-tuning	Accuracy	Macro Precision	Macro Recall	Macro F_1
AIDPATH	iBOT-Hist	Full	0.9957	0.9957	0.9958	0.9957
		LoRa	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		Linear Probe	0.9957	0.9958	0.9957	0.9957
	DINO-Hist	Full	0.9829	0.9829	0.9829	0.9829
		LoRa	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		Linear Probe	0.9957	0.9957	0.9958	0.9957
PathoSpotter-HE	iBOT-Hist	Full	0.8303	0.7903	0.7854	0.7871
		LoRa	**0.9051**	**0.8787**	**0.8785**	**0.8785**
		Linear Probe	0.7818	0.7543	0.7309	0.7409
	DINO-Hist	Full	0.6242	0.5836	0.5635	0.5717
		LoRa	0.8848	0.8631	0.8456	0.8532
		Linear Probe	0.6747	0.6622	0.6572	0.6423
PathoSpotter-PAS	iBOT-Hist	Full	0.7615	0.7292	0.7119	0.7181
		LoRa	**0.8619**	**0.8117**	0.8045	0.8077
		Linear Probe	0.7699	0.7273	0.7362	0.7262
	DINO-Hist	Full	0.5858	0.5965	0.5633	0.5742
		LoRa	0.8452	0.8102	**0.8288**	**0.8186**
		Linear Probe	0.6695	0.6541	0.6542	0.6523
PathoSpotter MultiStain	iBOT-Hist	Full	0.7996	0.7500	0.7431	0.7461
		LoRa	0.8497	0.8162	0.8035	0.8091
		Linear Probe	0.6935	0.6721	0.6658	0.6661
	DINO-Hist	Full	0.6238	0.6014	0.5857	0.5915
		LoRa	**0.8536**	**0.8272**	**0.8175**	**0.8215**
		Linear Probe	0.6601	0.6341	0.6466	0.6306
RCC	iBOT-Hist	Full	0.8592	0.869	0.8598	0.8499
		LoRa	**0.9225**	0.9116	0.9121	0.9097
		Linear Probe	0.9085	0.8954	0.8943	0.8929
	DINO-Hist	Full	0.8662	0.8522	0.8457	0.8464
		LoRa	**0.9225**	**0.9180**	**0.9169**	**0.9121**
		Linear Probe	0.8732	0.8716	0.8597	0.8606

4.2. CBIR Effectiveness

To evaluate the effectiveness of the models on the retrieval task, 50 images were randomly selected from the test set. For this scenario, we selected dataset PathoSpotter-HE as the reference. Table 4 presents the effectiveness of the iBOT-Hist and DINO-Hist models, considering different ranking depth levels ($MAP@5$, $MAP@10$, $MAP@20$, $MAP@30$, $MAP@40$, $MAP@50$). The results presented in Table 4 indicate that the DINO-Hist model outperforms iBOT-Hist at several ranking levels. Specifically, DINO-Hist demonstrates a statistically significant superiority at the $MAP@10$ and $MAP@40$ values, highlighted in bold. For the other levels, although DINO-Hist also presents higher values, the differences are not significant enough to be considered superior, indicating equivalence in effectiveness.

The superiority of DINO-Hist at several ranking levels, especially at $MAP@10$ and $MAP@40$, indicates that this model is better able to capture and recover subtle features of tissue images, a crucial aspect in identifying specific pathological patterns. This capability not only enhances diagnostic efficiency but can also reduce analysis time,

Table 3. Percentage gain of the LoRa technique compared to the Full and Linear Probe fine-tuning applied to the PathoSpotter-HE dataset

Model	Fine-tuning	Accuracy	Macro Precision	Macro Recall	Macro F_1
iBOT-Hist	LoRa	0.9051	0.8787	0.8785	0.8785
	Full	0.8303	0.7903	0.7854	0.7871
	Gain(%)	**9.01**	**11.10**	**11.80**	**11.64**
	LoRa	0.9051	0.8787	0.8785	0.8785
	Linear Probe	0.7818	0.7543	0.7309	0.7409
	Gain(%)	**15.78**	**16.42**	**20.25**	**18.56**
DINO-Hist	LoRa	0.8848	0.8631	0.8456	0.8532
	Full	0.6242	0.5836	0.5635	0.5717
	Gain(%)	**41.72**	**47.96**	**50.05**	**49.24**
	LoRa	0.8848	0.8631	0.8456	0.8532
	Linear Probe	0.6747	0.6622	0.6572	0.6423
	Gain(%)	**31.14**	**30.30**	**28.74**	**32.80**

Table 4. Effectiveness of the models considering the CBIR task. Statistical superiority is highlighted in bold; the others are considered equivalent.

Modelo	MAP@5	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50
iBOT-Hist	0.8550	0.7770	0.7128	0.6918	0.6746	0.6719
DINO-Hist	0.9070	**0.8756**	0.7971	0.7731	**0.7744**	0.7690

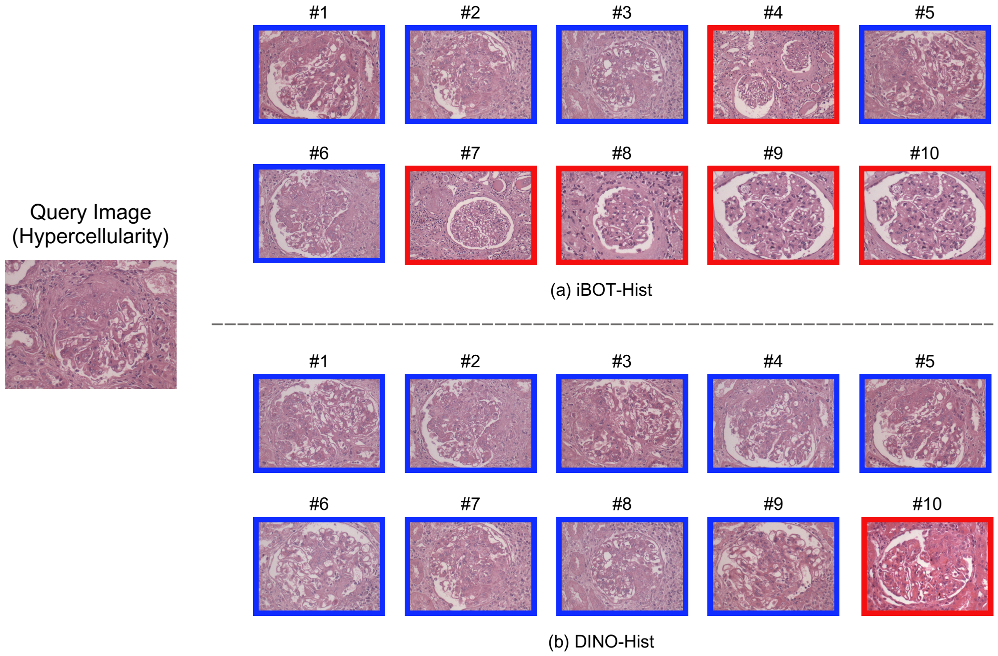
minimize diagnostic errors, and, consequently, improve patient care. These results reinforce the importance of using advanced machine learning models in medical applications, where efficiency and reliability are essential. Furthermore, they demonstrate the potential of self-supervised models, even those pre-trained with natural images, to adapt to other domains.

Figure 2 presents an example of a real search performed in the dataset PathoSpotter-HE. The image of the specified query belongs to the hypercellularity class. The first ten images retrieved by the system are presented. As can be seen, the most effective model was DINO-Hist. Of the ten images retrieved, only one does not belong to the class of interest. On the other hand, for iBOT-Hist, of the ten images, five are irrelevant to the query.

Overall, the results achieved were effective, demonstrating the potential of self-supervised architectures. This result is of great importance, especially considering the context in which this work is set. Indeed, in clinical practice, the ability to quickly identify similar histopathological images from large databases can assist pathologists in case comparison, accurate diagnosis and informed decision making.

5. Conclusion

In this study, we develop and experimentally evaluate machine learning models, pre-trained through self-supervised techniques, applied to both disease/injury classification and content-based image retrieval tasks. The results indicate that the investigated self-



(a) iBOT-Hist

(b) DINO-Hist

Figure 2. Top-10 results for the specified query image. At the top we have the result of iBOT-Hist (MAP@10 = 0.8412). At the bottom the result of DINO-Hist (MAP@10 = 0.8412). Relevant images are highlighted in blue, while irrelevant ones are highlighted in red.

supervised approaches achieved promising results when adapted to other contexts. The models were effective on several datasets, considering broad evaluation measures. The results demonstrated that the baseline DINOv2 model was numerically superior to iBOT. This result was driven by the application of the LoRa adaptation technique. When disregarding this strategy, iBOT outperforms DINOv2 on all datasets used in this study. This highlights the critical role of employing effective fine-tuning techniques, as the selected approach directly influences the model's effectiveness and generalization capability.

Regarding content-based image retrieval, the results showed that both architectures are effective. Considering the PathoSpotter-HE dataset, it was found that DINO-Hist was superior to iBOT-Hist, demonstrating statistical superiority at the *ranking* $MAP@10$ and $MAP@40$ levels. The experiments also indicated that the *fine-tuning* methodology employed directly influences the effectiveness of the models and, consequently, their adaptation to a new domain. The experiments demonstrated that the LoRa technique was more effective than the other variants. LoRa better adapted the base models to the context of histopathological images. This result was even more expressive in DINO-Hist, especially when compared to the Full fine-tuning, achieving gains above 40%. Despite the significant gains when applying LoRa, it is necessary to carry out new experiments to evaluate the influence of the dataset size on this result.

Acknowledgement

This work was partially supported by UEFS-AUXPPG 2025 and CAPES-PROAP 2025 grants. We also acknowledge the PathoSpotter project for contributing three of the five databases used in this study.

References

Abels, E. et al. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J Pathol*, 249(3):286–294.

Baharoon, M. et al. (2024). Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks.

Balestriero, R. et al. (2023). A cookbook of self-supervised learning.

Bueno, G. et al. (2020). Data for glomeruli characterization in histopathological images. *Data in Brief*, 29:105314.

Cerqueira, S. et al. (2021). Pathospotter classifier: Uma serviço web para auxílio à classificação de lesões em glomérulos renais. In *Anais do XXI SBCAS*, pages 60–70, Porto Alegre, RS, Brasil. SBC.

Chagas, P. et al. (2020). Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artif Intell Med*, 103:101808.

Chen, L. et al. (2019). Self-supervised learning for mia using image context restoration. *Med Image Anal*, 58:101539.

Filiot, A. et al. (2023). Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*.

Kataria, T. et al. (2023). To pretrain or not to pretrain? a case study of domain-specific pretraining for semantic segmentation in histopathology. In *Medical Image Learning with Limited and Noisy Data*, pages 246–256, Cham. Springer Nature Switzerland.

L'Imperio, V. et al. (2021). Digital pathology for the routine diagnosis of renal diseases: a standard model. *Journal of Nephrology*, 34(3):681–688.

Majumdar, S. et al. (2023). Gamma function based ensemble of cnn models for breast cancer detection in histopathology images. *Expert Systems with Applications*, 213:119022.

Mohammad Alizadeh, S. et al. (2023). A novel siamese deep hashing model for histopathology image retrieval. *Expert Systems with Applications*, 225:120169.

Oquab, M. et al. (2023). Dinov2: Learning robust visual features without supervision.

Ozen, Y. et al. (2021). Self-supervised learning with graph neural networks for region of interest retrieval in histopathology. In *ICPR*, pages 6329–6334.

Wickstrøm, K. K. et al. (2023). A clinically motivated self-supervised approach for content-based image retrieval of ct liver images. *CMIG*, 107:102239.

Yamaguchi, R. et al. (2021). Glomerular classification using convolutional neural networks based on defined annotation criteria and concordance evaluation among clinicians. *Kidney International Reports*, 6(3):716–726.

Yang, P. et al. (2020). A deep metric learning approach for histopathological image retrieval. *Methods*, 179:14–25. Interpretable machine learning in bioinformatics.

Zheng, Y. et al. (2022). Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval. *Med Image Anal*, 76:102308.