Boosting Vision-Language Models for Histopathology Classification: Predict all at once

Maxime Zanella[1,2⋆], Fereshteh Shakeri[3,4⋆], Yunshi Huang[3,4], Houda Bahig[4], and Ismail Ben Ayed[3,4]

[1] Université Catholique de Louvain (UCLouvain), Louvain-La-Neuve, Belgium
maxime.zanella@uclouvain.be
[2] Université de Mons (UMons), Mons, Belgium
[3] École de Technologie Supérieure (ÉTS), Montréal, Canada
fereshteh.shakeri.1@etsmtl.net
[4] Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Canada

Abstract. The development of vision-language models (VLMs) for histopathology has shown promising new usages and zero-shot performances. However, current approaches, which decompose large slides into smaller patches, focus solely on inductive classification, *i.e.,* prediction for each patch is made independently of the other patches in the target test data. We extend the capability of these large models by introducing a transductive approach. By using text-based predictions and affinity relationships among patches, our approach leverages the strong zero-shot capabilities of these new VLMs without any additional labels. Our experiments cover four histopathology datasets and five different VLMs. Operating solely in the embedding space (*i.e.,* in a black-box setting), our approach is highly efficient, processing 10^5 patches in just a few seconds, and shows significant accuracy improvements over inductive zero-shot classification. Code available at https://github.com/FereshteShakeri/Histo-TransCLIP.

Keywords: Histopathology · Medical VLMs · Zero-Shot Learning · Transductive Inference · Efficient Adaptation

1 Introduction

Histology slides obtained from Whole Slide Image (WSI) [**18**] scanners play a crucial role in cancer diagnosis and staging [16]. These slides offer a detailed view of diseased tissues, aiding in the determination of treatment options. Pathologists primarily diagnose cancers by examining WSIs to identify different tissue types. However, manually analyzing these WSIs imposes a significant workload, leading to substantial delays in reporting time. Moreover, in real clinical environments, the classification of cancer-related tissues is highly diverse, encompassing various cancer sites. Even within a single cancer site, tasks can vary in their levels of class granularity. Therefore, automating tissue-type classification in histology

⋆ M. Zanella and F. Shakeri—Equal contribution.

images ([1, 12, 19, 20, 25] to list a few) holds significant clinical value but is hindered by the difficulty of collecting large labeled datasets and the variability of fine-grained labels.

The advent of multi-modal learning methods that process and integrate information from diverse modalities has alleviated some issues of training fine-grained classifiers and collecting costly labeled data. In particular, vision-language models (VLMs) such as CLIP [21] and ALIGN [9] have gained popularity in computer vision, and demonstrated promising generalization capabilities across various downstream tasks. These so-called foundation models jointly train vision and text embeddings using contrastive learning on large-scale image-text datasets. This new multi-modal paradigm can naturally be extended to clinical scenarios, where combinations of multiple data modalities–mainly texts and images–are often adopted to obtain more accurate and comprehensive diagnosis. For example, clinical notes and pathology reports, alongside histopathology slides, are commonly used for throughout analysis [6]. However, the direct application of deep learning techniques, more specifically vision-language pre-training strategies, to medical imaging is complex, due to the lack of fine-grained expert medical knowledge, which is required to capture specialized information [4]. This issue has been partly addressed for histopathology slides by collecting diverse data from scientific publications, Twitter, or even YouTube videos [7, 8, 15].

Current usage of such models predominantly align with the *inductive* paradigm, i.e., inference for each test sample is performed independently from the other samples within the test dataset. In contrast, *transduction* performs joint inference on all the test samples of a task, leveraging the statistics of the target unlabeled data [10, 26]. Transduction has primarily been explored for few-shot classification of natural images, to tackle the inherent challenges of training under limited supervision [3, 5]. These techniques utilize labeled samples to transfer information to unlabeled test data. Interestingly, in the novel multi-modal paradigm introduced by VLMs, supervision can be instead provided through textual descriptions of each classes (prompts), in a zero-shot setting, *e.g.,* `a pathology tissue showing` [`class name`]. Along with their corresponding representation derived from the language encoder, similarities between text and image embeddings can be leveraged to enable transductive inference even in the *zero-shot* scenario, as pointed by recent works in computer vision [17, 29] (see Figure 1).

Contributions. With the ongoing development of foundation models in medical imaging and specifically histopathology, and the potential application of transductive inference, our objective is to improve zero-shot predictions of VLMs within this framework. Our main contributions can be summarized as follows:

- We compare the zero-shot performance of vision-language models for histology and propose an effective transductive method to significantly boost their accuracy by leveraging the structure among patches during inference.
- Our transductive approach does not require labels; instead, it utilizes text-based predictions as regularization.

– To alleviate the computational workload, our method relies on the pre-computed features only, without access to the pre-trained weights, thus accommodating black-box constraints. This makes it feasible to process very large-scale slides in a matter of seconds.

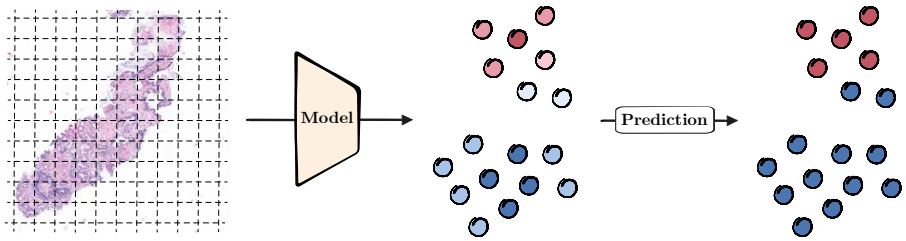
2 Related Work

VLMs for histology. Unlike natural images, which are often available in millions (*e.g.,* CLIP [21] is trained on 400M image-text pairs), clinical image-text pairs are more challenging to amass. Similar to other works introducing VLMs for medical imaging (*e.g.,* for radiology [27, 28, 30], or ophthalmology [23]), several VLMs for computational pathology has appeared recently, differentiating themselves primarily through their data collection and curation methodologies. PLIP [7] curates OpenPath, a large dataset of pathology images paired with text descriptions. Quilt-1M [8] stands as one of the largest vision-language histopathology dataset to date, comprising 1 million image/text pairs sourced from YouTube videos. More recently CONCH [15] integrates parts of the PubMed Central Open Access Dataset yielding 1.17 million samples. As these new VLMs have been developed in a short amount of time, determining the most suitable one is not straightforward. Therefore, we provide a comparison of these models and demonstrate the applicability of our approach across each of them.

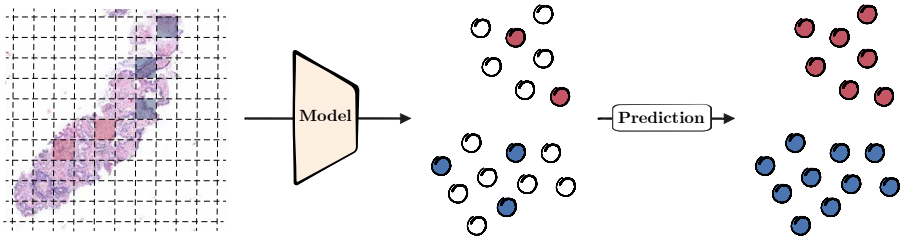
Transductive learning. In the few-shot literature solely based on vision models, transduction leverages both the few labeled samples and unlabeled test data [10, 26], outperforming inductive methods [3, 5, 14, 31]. This setting, widely explored in computer vision, has been recently deployed in histopathology [22], using the annotations of a few patches from slides of liver. However, previously mentioned transductive methods have been shown to suffer from significant performance drops when applied to VLMs [17, 29]. This motivated a few, very recent transductive methods in computer vision, focusing on natural images and explicitly leveraging the textual modality along the image embeddings [17, 29]. In contrast to [22], our work exploits the findings and transductive-inference zero-shot objective in [29], aiming to boost the predictive accuracy of pretrained histopathology VLMs without any supervision.

3 Method

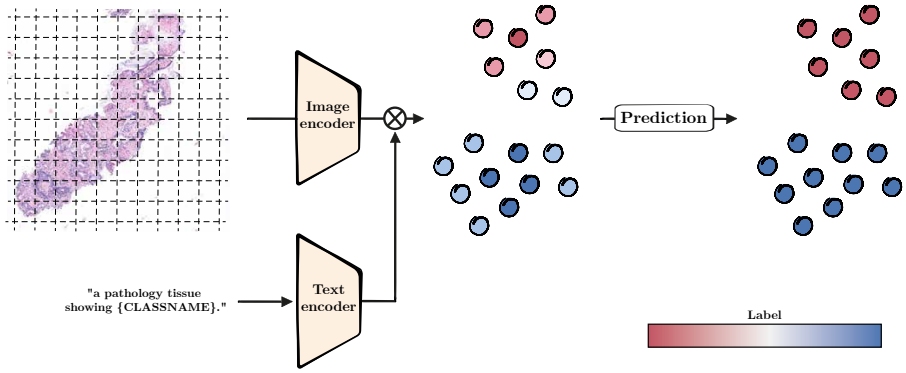
In this section, we describe the Histo-TransCLIP objective function for transductive inference in VLMs, for the K-class prediction problem. This objective function depends on two types of variables: (i) assignment variables $\mathbf{z}_i = (z_{i,k})_{1 \leq k \leq K} \in \Delta_K$, for each patch $i \in \mathcal{Q}$; and (ii) Gaussian Mixture Model (GMM) parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_{1 \leq k \leq K}$ and $\boldsymbol{\Sigma}$. We will first detail the main components of Histo-TransCLIP, before deriving the overall procedure.



(a) In the typical inductive setting, a model is trained and then used to infer on each patch separately. This approach can be efficient when large annotated datasets for each task are available. This procedure often involves predicting the most probable class (argmax).



(b) In the traditional transductive few-shot setting, a pre-trained encoder (e.g., on ImageNet or large-scale histology dataset) requires manual annotations for the new task to propagate information from labeled to unlabeled samples. This process often involves measuring affinities or distances between encoded samples.



(c) VLMs leverage textual descriptions of each class to generate pseudo-labels without any manual annotation. These initial predictions can then be refined, for example, by leveraging the data structure.

Fig. 1: Illustration depicting histopathology classification in the inductive setting (a), the commonly-used few-shot transductive setting (b), and the zero-shot transductive setting enabled by VLMs (c).

Gaussian modelization We modelize the likelihood of the target data as a balanced mixture of multivariate Gaussian distributions, each representing a class k, parameterized by mean vector $\boldsymbol{\mu}_k$ and a diagonal (shared among classes) covariance matrix $\boldsymbol{\Sigma}$:

$$p_{i,k} \propto \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}_i - \boldsymbol{\mu}_k)\right)$$

where $\mathbf{f}_i$ represents the encoding of patch i.

Text-based predictions. When dealing with a zero-shot classification problem based on a VLM, and given a set of K candidate classes, one can get textual embeddings $\mathbf{t}_k$ (*e.g.,* from `a pathology tissue showing` [`kth class name`], $k = 1, \ldots, K$). Then pseudo-labels can be obtained by evaluating the softmax function of the cosine similarities between these two encoded modalities with τ being a temperature parameter:

$$\hat{y}_{i,k} = \frac{\exp(\tau \mathbf{f}_i^\top \mathbf{t}_k)}{\sum_j \exp(\tau \mathbf{f}_i^\top \mathbf{t}_j)} \tag{1}$$

Laplacian regularization. Laplacian regularizers are widely used in the context of graph/spectral clustering. This term encourages related samples (*i.e.,* pairs of patches with high affinity $w_{i,j}$) to have similar label assignments. We build affinities based on the cosine similarities of each patch representation:

$$w_{ij} = \mathbf{f}_i^\top \mathbf{f}_j \tag{2}$$

In fact, affinity relations can be modified for each specific use-case, allowing to inject knowledge in the optimization process. In our case, we can leverage the strong embedding capabilities of the image encoder to regularize the transductive procedure. In practice, to reduce memory needs, we sparsify the matrix by retaining only the 3 nearest neighbors of each patch.

Objective function. We minimize the following objective:

$$\mathcal{L}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i \in \mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \underbrace{-\sum_{i \in \mathcal{Q}}\sum_{j \in \mathcal{Q}} w_{ij}\mathbf{z}_i^\top \mathbf{z}_j}_{\text{Laplacian regularization}} + \underbrace{\sum_{i \in \mathcal{Q}} \mathrm{KL}(\mathbf{z}_i || \hat{\mathbf{y}}_i)}_{\text{Prediction penalty}} \tag{3}$$

The Kullback–Leibler (KL) term encourages the prediction $\mathbf{z}_i$ not to deviate significantly from the zero-shot prediction $\hat{\mathbf{y}}_i$, thereby providing text supervision without the need of any labels.

Procedure. We refer to [29] for the technical details about the derivation. Optimizing (3), subject to simplex constraints, we obtain the following decoupled

Pseudocode 1: Histo-TransCLIP procedure for transductive inference alternates between assignments and GMM-parameters updates.
Input: $\mathbf{f}$ are the image embeddings, $\mathbf{t}$ are the text/class embeddings, τ is the temperature scaling used during each VLM pretraining.

function `Histo-TransCLIP(f, t, τ)`
  `// Text-based pseudo-labels ŷ`
  $\hat{\mathbf{y}}_i = \text{softmax}(\tau \mathbf{f}_i^T \mathbf{t}) \quad \forall i$
  `// Initialize z, μ, Σ`
  $\mathbf{z}_i = \hat{\mathbf{y}}_i \quad \forall i$
  $\boldsymbol{\mu}_k = \text{top\_confident\_average}(\mathbf{f}, \hat{\mathbf{y}}) \quad \forall k$
  $\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{n\_features}$
  `// Iterative procedure`
  while `not_converged` **do**
    for `l = 1:...` **do**
      $\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \quad \forall i$
    $\boldsymbol{\mu}_k = \frac{\sum_{i \in \mathcal{Q}} z_{i,k} \mathbf{f}_i}{\sum_{i \in \mathcal{Q}} z_{i,k}} \quad \forall k$
    $\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2$
  return `z`

update rules for the assignment variables, which can be computed in parallel for all samples (*i.e.,* patches) at a given iteration l:

$$\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{D}} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{D}} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \tag{4}$$

Note how each assignment $\mathbf{z}_i$ depends on its neighbors. This update must be computed iteratively until convergence, enabling assignments to propagate from the GMM likelihood to neighboring samples, weighted by their affinity. Since these updates are decoupled, this step can be parallelized efficiently (see runtime in Table 2). With other variables fixed, we then have the following closed-form updates for the GMM parameters:

$$\boldsymbol{\mu}_k = \frac{\sum_{i \in \mathcal{Q}} z_{i,k} \mathbf{f}_i}{\sum_{i \in \mathcal{Q}} z_{i,k}} \tag{5}$$

$$\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2 \tag{6}$$

The procedure is summarized in Pseudocode 1 and alternates between solving (4) to get assignments for each patch and computing the GMM parameters (5) (6) according to those assignments until convergence (see proof in [29]).

Table 1: Zero-shot and Histo-TransCLIP performance on top of various VLMs. Best values are highlighted in **bold**. $\Delta_{\textbf{transductive}}$ is the average accuracy gain brought by our transductive approach.

Dataset	Method	Model				
		CLIP	Quilt-B16	Quilt-B32	PLIP	CONCH
SICAP-MIL	Zero-shot	**29.85**	40.44	**35.04**	46.84	27.71
	Histo-TransCLIP	24.72	**58.49**	28.18	**53.23**	**32.58**
LC(Lung)	Zero-shot	**31.46**	43.00	76.24	84.96	84.81
	Histo-TransCLIP	25.62	**50.53**	**93.93**	**93.80**	**96.29**
SKINCANCER	Zero-shot	4.20	15.38	39.71	22.90	58.53
	Histo-TransCLIP	**11.46**	**33.33**	**48.80**	**36.72**	**66.22**
NCT-CRC	Zero-shot	25.39	29.61	53.73	63.17	66.27
	Histo-TransCLIP	**39.61**	**48.40**	**58.13**	**77.53**	**70.36**
Average	Zero-shot	22.73	32.1	51.18	54.47	59.33
	Histo-TransCLIP	**25.35**	**47.69**	**57.26**	**65.32**	**66.36**
	$\Delta_{\textbf{transductive}}$	**+2.62**	**+15.59**	**+6.08**	**+10.85**	**+7.03**

4 Experiments

We conduct a comprehensive comparison of several vision-language models pretrained on histology images, namely PLIP [7], QUILT [8] (for which we report two versions) and CONCH [15]. Text embeddings for each category are obtained following the specific 22 prompts used for CONCH (only one name is assigned to each target class), which are then averaged to get a single textual embedding per class. Numerical results are top-1 accuracy which compare zero-shot prediction (*i.e.,* inductive inference) and Histo-TransCLIP (*i.e.,* transductive inference).

Datasets. We study different histopathology classification tasks on various organs and cancer types [2, 11, 13, 24]. Specifically, *NCT-CRC* [11] comprises patches of colorectal adenocarcinoma categorized into 9 classes, *SICAP-MIL* [24] includes 4 prostate cancer grading, *SKINCANCER* [13] is annotated with 9 anatomical tissue structures, and *LC25000(Lung)* [2] focuses on 3 classes of lung cancer. These diverse benchmarks enable the study of the generalization capability of VLMs pretrained on histology images and provide a thorough assessment of our transductive approach.

Results. Table 1 presents a comparative analysis of zero-shot performance and the improvement achieved by Histo-TransCLIP. The lower classification accuracy of CLIP emphasizes the need for VLMs specifically pretrained on histology. Notably, the recently proposed CONCH model demonstrates the highest average accuracy. Note that the variation in zero-shot accuracies compared to the

Table 2: Features denotes the runtime to pre-compute the image and text embeddings, Histo-TransCLIP denotes the runtime of our transductive procedure once embeddings are provided. Experiments were conducted on a single NVIDIA GeForce RTX 3090 (24Gb) GPU.

#Patches	Features	Histo-TransCLIP
10^2	$\sim$ 1 sec.	$\sim$ 0.1 sec.
10^3	$\sim$ 4 sec.	$\sim$ 0.2 sec.
10^4	$\sim$ 28 sec.	$\sim$ 0.4 sec.
10^5	$\sim$ 5 min.	$\sim$ 6 sec.

original paper values is largely influenced by the choice of prompt templates, for instance PLIP zero-shot results are significantly improved. This yields interesting questions on prompt sensitivy as discussed for future work in Section 5. Histo-TransCLIP consistently enhances performance significantly, highlighting the benefits of its transductive approach. Only in a few cases does the accuracy of Histo-TransCLIP drop, particularly when zero-shot performance is low due to direct regularization with initial text predictions. In most cases, Histo-TransCLIP effectively enhances performance, even on tasks initially achieving high accuracy, showcasing its strong ability to refine slightly misaligned text predictions for various VLMs.

Computational workload. Table 2 details the computational overhead associated with Quilt-B16 visual and textual feature extraction, alongside the implementation of Histo-TransCLIP across varying patch numbers in the *NCT-CRC* dataset. While the time for feature extraction increases with the number of patches, the additional workload introduced by Histo-TransCLIP remains negligible. This shows transduction can importantly improve performance while maintaining black-box adaptation (*i.e.,* without accessing the model's parameters) and without adding any notable additional workload.

5 Conclusion

We have demonstrated the significant value that transduction can bring to histology. By leveraging text-based predictions through a Kullback–Leibler divergence penalty and incorporating shared information among patches with Laplacian regularization, our approach significantly enhances the performance of vision-language models. Notably, our method is highly efficient and does not require additional labels or access to model parameters.

Future Work. Our approach can be naturally extended to the few-shot setting. Additionally, the quality of the prompts, *i.e.,* the textual descriptions of each class, can significantly impact the final zero-shot performance. Studying this impact is undoubtedly valuable for safer applications. Finally, while our current

work focuses on transduction using patches from multiple slides, a more constrained and valuable application would involve transduction on patches from a single slide to improve performance on a per-patient basis.

Acknowledgments. M. Zanella is funded by the Walloon region under grant No. 2010235 (ARIAC by DIGITALWALLONIA4.AI). F. Shakeri is funded by Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Bilgin, C., Demir, C., Nagi, C., Yener, B.: Cell-graph mining for breast tissue modeling and classification. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5311–5314. IEEE (2007)
2. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
3. Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P., Ben Ayed, I.: Information maximization for few-shot learning. Advances in Neural Information Processing Systems **33**, 2445–2457 (2020)
4. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79** (2022)
5. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: International Conference on Learning Representations (2019)
6. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. CoRR **abs/2403.02469** (2024). `https://doi.org/10.48550/ARXIV.2403.02469`, `https://doi.org/10.48550/arXiv.2403.02469`
7. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**, 1–10 (2023)
8. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. arXiv preprint arXiv:2306.11207 (2023)
9. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916 (2021)
10. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of the Sixteenth International Conference on Machine Learning. pp. 200–209 (1999)
11. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018)
12. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal **16**, 34–42 (2018)

13. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliss, R.R., Muley, T., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. Frontiers in Oncology **12**, 1022967 (2022)
14. Liu, J., Song, L., Qin, Y.: Prototype rectification for few-shot learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 741–756. Springer (2020)
15. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
16. Madabhushi, A.: Digital pathology image analysis: opportunities and challenges. Imaging in medicine **1**(1), 7 (2009)
17. Martin, S., Huang, Y., Shakeri, F., Pesquet, J.C., Ben Ayed, I.: Transductive zero-shot and few-shot clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28816–28826 (2024)
18. Pantanowitz, L.: Digital images and the future of digital pathology. Journal of pathology informatics **1** (2010)
19. Petushi, S., Garcia, F.U., Haber, M.M., Katsinis, C., Tozeren, A.: Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC medical imaging **6**(1), 1–11 (2006)
20. Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R., Gurcan, M.: Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 196–204. Springer (2008)
21. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
22. Sadraoui, A., Martin, S., Barbot, E., Laurent-Bellue, A., Pesquet, J.C., Guettier, C., Ben Ayed, I.: A transductive few-shot learning approach for classification of digital histopathological slides from liver cancer. In: IEEE International Symposium on Biomedical Imaging (ISBI) (2024)
23. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. ArXiv Preprint (2023)
24. Silva-Rodríguez, J., Schmidt, A., Sales, M.A., Molina, R., Naranjo, V.: Proportion constrained weakly supervised histopathology image classification. Computers in Biology and Medicine **147**, 105714 (2022)
25. Tabesh, A., Teverovskiy, M., Pang, H.Y., Kumar, V.P., Verbel, D., Kotsianti, A., Saidi, O.: Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE transactions on medical imaging **26**(10), 1366–1378 (2007)
26. Vapnik, V.: An overview of statistical learning theory. IEEE Transactions on Neural Networks **10**(5), 988–999 (1999). https://doi.org/10.1109/72.788640
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1–12 (10 2022)
28. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: ICCV (2023)
29. Zanella, M., Gérin, B., Ayed, I.B.: Boosting vision-language models with transduction. arXiv preprint arXiv:2406.01837 (2024)
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: MHLC (2022)

31. Ziko, I., Dolz, J., Granger, E., Ayed, I.B.: Laplacian regularized few-shot learning. In: International conference on machine learning. pp. 11660–11670. PMLR (2020)