The Self-Supervision Regime and Encoder Fit for Histopathology Image Analysis

Asfandyar Azhar AAZHAR@ANDREW.CMU.EDU
Carnegie Mellon University

Abstract

This study systematically compares vision transformers (ViTs) and residual neural networks (ResNets) for slide-level histopathology analysis using self-supervised learning (SSL) strategies SimCLR and DINO, alongside ImageNet-initialized counterparts. Binary and multiclass classification were evaluated on the TCGA-CRCk and PANDA datasets using the multiple instance learning paradigm. Grad-CAM and attention rollout were applied for visual explainability for ResNets and ViTs, respectively. The results showed that SimCLR excelled with smaller architectures, while DINO performed better with larger ones. ResNets underperformed with DINO regardless of architecture size, whereas ViTs excelled on the PANDA dataset but required scaling to perform well on TCGA-CRCk, a more complex dataset. An emerging *self-supervision regime and encoder fit* suggests that DINO-ViTs and SimCLR-ResNets are well-suited pairs, though the effects of SSL methods on encoders remain unclear. ViTs demonstrated superior scalability and interpretability compared to ResNets, but further research is needed to better understand the interplay between SSL methods and encoders.

Keywords: Computer Vision, Self-Supervised Learning, Histopathology, Multiple Instance Learning, Explainability

1. Introduction

The advent of deep learning has revolutionized histopathology by enabling the analysis of complex biomedical image data with unprecedented scale and precision. Among the leading deep learning architectures, vision transformers (ViTs) (Dosovitskiy et al., 2021) and residual neural networks (ResNets) (He et al., 2015) have achieved remarkable results on natural image benchmarks, such as ImageNet (Deng et al., 2009). Notably, ViTs have outperformed ResNets on several tasks, suggesting potential advantages in domains requiring nuanced feature extraction. However, the efficacy of ViTs over ResNets in histopathology image analysis remains underexplored, with limited studies yielding mixed results (Chen and Krishnan, 2022). This study aims to systematically investigate the performance of ViTs and ResNets in histopathology, specifically in slide-level classification tasks. We evaluate these models under different pretraining strategies, leveraging self-supervised learning (SSL) methods, SimCLR (Chen et al., 2020) and DINO (Caron et al., 2021), and compare them against ImageNet-initialized counterparts. The goal is to determine whether SSL-pretrained ViTs can offer a significant performance improvement over traditional ResNet models in this domain.

To address this gap, we implement a multiple instance learning (MIL) framework (Ilse et al., 2018), where each whole slide image (WSI) is processed as a collection of patches. ViTs encode each patch independently, aggregating the resulting class (CLS) tokens to predict the slide-level label. We conduct a series of experiments comparing both ViTs and ResNets, pretrained with SimCLR and DINO, as well as ImageNet-initialized models. To ensure the robustness and generalizability of our findings, we use two different datasets: TCGA-CRCk (Schirris et al., 2022) for binary classification (benign vs. malignant) and PANDA (challenge consortium, 2022) for multiclass classification (cancer grading). These datasets allow us to evaluate model performance across different classification tasks, providing a comprehensive understanding of each model's strengths and weaknesses. Our research addresses two critical questions:

1. *Do ViTs outperform ResNets in classifying histopathology images at the slide level?*

2. *How does the choice of SSL method affect the performance of these models in slide-level classification tasks?*

By answering these questions, we aim to determine the most effective deep learning architecture and pretraining strategy for histopathology image analysis. This work could lead to more accurate diagnostic tools, ultimately improving patient outcomes by enabling better decision-making based on computational pathology.

Additionally, we will apply explainable AI (XAI) techniques, including GradCAM (Selvaraju et al., 2019) for ResNets and attention rollout (Abnar and Zuidema, 2020) for ViTs, to visualize the regions deemed important by these models. This analysis could uncover novel pathological features and enhance our understanding of disease processes at a microscopic level, providing new insights into the complex patterns of histopathological data.

In histopathology, distinguishing between slide-level and patch-level classification is crucial. Slide-level classification involves assigning a single label to an entire WSI, which may indicate whether a slide contains benign or malignant tissue, or classify it into specific categories based on disease severity. Patch-level classification, on the other hand, breaks down WSIs into smaller, more manageable patches, each classified independently. This approach can identify specific regions of interest (ROIs) within the tissue, such as areas with high cell density or malignancy. Our study focuses on slide-level classification due to its efficiency and lower computational demand, providing a broad overview of the tissue. However, this approach may miss fine details that patch-level classification could capture. Understanding these distinctions and their implications for deep learning applications is critical for developing more accurate and efficient diagnostic models in computational pathology.

2. Related Work

Deep Multiple Instance Learning. Deep multiple instance learning (MIL) has become a pivotal approach for histopathology image analysis, addressing the challenge of learning from weakly labeled data. Ilse et al. (2018) proposed an attention-based deep MIL framework, where the MIL problem is modeled to maximize a log-likelihood function under the assumption that bag labels follow a Bernoulli distribution. To maintain permutation invariance in bag probability, $\Theta(X)$, the approach applies a three-step process: transformation of instances to embeddings, permutation-invariant aggregation (MIL pooling), and a final transformation for bag probability estimation.

The foundational theory behind this method leverages the properties of symmetric functions, as described by Zaheer et al. (2018) and Qi et al. (2017), allowing any set function to be decomposed into these steps. Theorems from those works provide the mathematical basis for modeling $\Theta(X)$ using functions like sum or max pooling to achieve permutation invariance, crucial for learning meaningful bag-level features.

Two primary approaches exist for deep MIL: the instance-based approach, where each instance is scored independently before aggregation, and the embedding-based approach, which transforms instances into a lower-dimensional embedding before pooling. While the instance-based approach is interpretable, the embedding-based approach often offers better performance by capturing more complex relationships within the data (Liu et al., 2012). However, embedding-based methods lack interpretability, a gap that attention-based MIL pooling, as introduced by Ilse et al. (2018), aims to fill by assigning importance weights to instances based on their contribution to the bag label.

Attention-based MIL pooling models, using mechanisms like soft-attention layers, provide a flexible and adaptable framework for MIL, potentially outperforming traditional pooling methods. The use of attention allows for the model to focus on relevant instances within a bag, improving the interpretability and accuracy of bag-level predictions (Wang et al., 2018). However, this approach still assumes independence among instances, which may not always hold in practice.

Modeling Dependencies with Self-Attention. To address the independence assumption limitation, the self-attention mechanism, originally developed for capturing long-range dependencies in sequential data (Vaswani et al., 2017), has been adapted for MIL tasks. Self-attention models allow for the consideration of dependencies among instances within a bag, enabling more nuanced and accurate modeling of bag-level features. The mechanism computes weighted sums of input vectors, where weights are learned based on the relevance of each instance to others, thus capturing dependencies directly. This process enhances the model's ability to understand the relationships between different instances, crucial for complex histopathology tasks where spatial and contextual information is vital. The incorporation

of self-attention in MIL models has been shown to outperform traditional models, especially in scenarios where instance independence is a strong assumption (Rymarczyk et al., 2021).

Recent studies have further refined these approaches, incorporating spatial hierarchies and positional encodings to capture the structure of histopathology images better. This method preserves spatial context and enables the model to learn more robust features that are essential for accurate pathology diagnosis (Xiong et al., 2023; Chen et al., 2022).

SSL in Histopathology. The application of SSL methods, particularly with ViTs, has shown promise in histopathology image analysis. Chen and Krishnan (2022) demonstrated that SSL, specifically using DINO-based knowledge distillation, enables ViTs to learn meaningful morphological features from pathology data without requiring extensive labeled datasets. Their work highlighted that ViTs could leverage SSL to identify distinct morphological phenotypes, thus showing potential in understanding histopathological patterns at a deeper level. Their findings indicated that ViTs pretrained with SSL methods outperformed conventional ResNet-50 models pretrained on ImageNet for histopathology tasks such as tissue classification and cancer grading. Notably, DINO's attention mechanisms helped localize critical tissue structures, suggesting that ViTs could effectively capture important inductive biases in histopathology. While their approach established a strong baseline, it primarily focused on ViTs with DINO pretraining, leaving room for exploring other SSL methods and backbone architectures like ResNets. This study aims to extend these findings by systematically evaluating the impact of different SSL strategies and architectures on histopathology tasks, thereby providing a comprehensive understanding of the strengths and limitations of each approach in a clinical context.

3. Methodology

3.1. Problem Formulation

Pathology involves the scientific study of diseases through various methodologies, including biopsies, where tissue samples are analyzed to diagnose medical conditions. In cancer pathology, histology WSIs are crucial for determining clinical endpoints like cancer subtype, grade, and stage (Amin et al., 2017). Despite being the gold standard, manual WSI analysis is subjective and prone to significant inter- and intra-observer variability, especially in tasks like Gleason scoring for prostate cancer (Nicholson et al., 2001; Carlson et al., 1998).

WSIs are typically gigapixel-sized images, reaching up to $150,000 \times 150,000$ pixels at $20\times$ magnification. This high resolution necessitates specialized viewers for detailed examination and annotation of regions of interest (ROIs). The massive size of these images makes direct application of state-of-the-art image classification models computationally infeasible due to their resource requirements, which exceed the capabilities of typical hardware setups (Tizhoosh and Pantanowitz, 2018). Moreover, the lack of annotated data presents another challenge. Annotating ROIs on WSIs is labor-intensive, requiring extensive time, focus, and domain expertise. Current deep learning models, such as CNNs, often require pixel-level annotations, which are impractical for large datasets due to high annotation costs and the required domain expertise. This limitation restricts the development of comprehensive phenotyping algorithms for histopathology.

To overcome these challenges, Multiple Instance Learning (MIL) is employed. MIL enables weak supervision of WSIs using slide-level labels, bypassing the need for exhaustive annotations. The two-step MIL process involves extracting instance-level embeddings from non-overlapping tissue patches and aggregating these embeddings to form a WSI-level representation (Campanella et al., 2019). This approach reduces the computational burden and leverages the power of deep learning in pathology without requiring extensive manual annotation. However, a key limitation remains the lack of diverse, well-curated pathology datasets, which hinders generalization across different tissues and organs. As a workaround, pre-trained encoders, such as ResNet on ImageNet, are used as feature extractors. While effective, these models may not fully capture domain-specific features due to differences between natural and histopathological images (Lu et al., 2020).

SSL emerges as a promising alternative, allowing for feature learning without labeled data by using auxiliary tasks to maintain consistency in feature representations. However, SSL faces two major limitations: a lack of comprehensive benchmarks for slide-level tasks and limited introspective assessments of learned features. The main challenges addressed in our study are: (1) The high resolution of WSIs; (2) The scarcity of annotated data; (3) Limitations of ex-

isting SSL models; (4) The need for diverse pathology datasets. Our objective is to enhance computational pathology by developing models capable of classifying WSIs and highlighting ROIs. This would significantly aid pathologists by providing an auxiliary diagnostic perspective. We aim to address the first three challenges, using TCGA-CRCk and PANDA datasets for pretraining and evaluation. Our ultimate goal is to determine which architectures and SSL frameworks are most effective for histopathology image analysis, thus enriching the computational pathologist's AI toolbox.

3.2. Datasets

TCGA-CRCk. The TCGA-CRCk dataset (Schirris et al., 2022) consists of preprocessed colorectal tumor tiles derived from formalin-fixed, paraffin-embedded (FFPE) WSIs from The Cancer Genome Atlas (TCGA). The dataset is designed for classifying microsatellite instability (MSI) versus stability (MSS) in colorectal cancer—a binary classification task where MSI indicates genetic hypermutability due to impaired DNA mismatch repair. TCGA-CRCk includes 192,314 tiles from 360 patients, split into training (93,408 tiles from 260 patients) and testing (98,906 tiles from 100 patients) sets. The training set consists of 39 MSI patients and 221 MSS patients, while the testing set includes 26 MSI and 74 MSS patients. All tiles are approximately at 10× magnification and pre-tiled.

PANDA. The PANDA dataset (challenge consortium, 2022) is a comprehensive collection of around 11,000 high-resolution prostate biopsy WSIs. Each WSI is annotated with the International Society of Urological Pathology (ISUP) grade and Gleason score, which assess prostate cancer severity. The dataset's unique feature is its high-resolution images, sometimes exceeding a billion pixels, necessitating sophisticated image processing techniques. The slides are sourced from multiple medical centers and scanned at different resolutions, providing a diverse dataset. This diversity enhances model robustness and generalizability across various histopathological slides. The WSIs are treated as bags of patches in a MIL framework, and the dataset is pre-tiled at 5× magnification.

3.3. Pretraining with Self-Supervised Learning Frameworks

Pretraining on ImageNet (Baseline Solution). ImageNet is a large-scale image dataset designed for computer vision research, containing over 14 million annotated images across more than 20,000 categories (Deng et al., 2009). For this study, we use the ImageNet-1K subset, which comprises 1.28 million images from 1,000 categories, providing a rich source of diverse visual data for pretraining a baseline for comparison. ViTs pretrained on ImageNet (Deng et al., 2009) serve as a baseline to evaluate their effectiveness in histopathology. The hypothesis is that while ImageNet-pretrained models offer a general feature space, models pretrained on domain-specific histopathology images might perform better due to closer alignment with the downstream task's distribution (Farahani et al., 2020). We employ two SSL methods—SimCLR and DINO—for pretraining, comparing their effectiveness against the ImageNet baseline to assess the impact of domain-specific pretraining.

Pretraining Regime. We adopt SSL methods, SimCLR and DINO, for pretraining ViTs and ResNets on histopathology images from TCGA-CRCk and PANDA datasets. Figure 1 illustrates our pipeline.

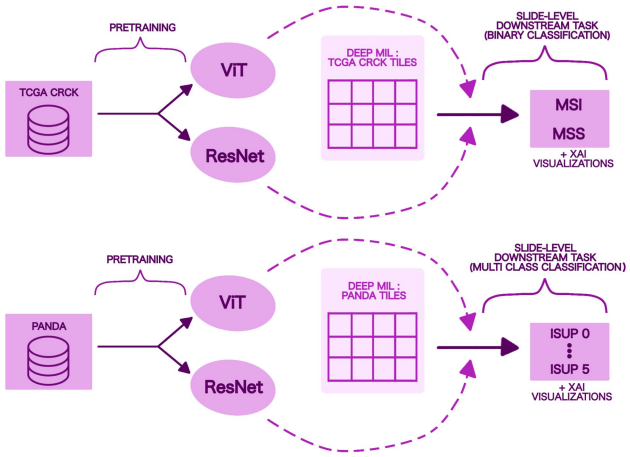

Figure 1: Experimental Design Flow Diagram.

Both SSL methods function by taking an image i and its augmented view i' and contrasting their representations $f(i)$ and $f(i')$. SimCLR aims to learn invariant representations by maximizing similarity between augmented views of the same image (positive samples) while minimizing similarity across different images (negative samples). In contrast, DINO employs a teacher-student paradigm, using a cross-entropy loss to match the output distri-

butions of a teacher network f_t and a student network f_s (Chen et al., 2022). DINO's advantage is its lack of reliance on negative samples, which is beneficial in datasets with significant class imbalance, common in histopathology. SimCLR uses data augmentations like color jittering, random crops, and flips, whereas DINO also employs local and global crops to foster correspondences, leveraging the part-whole hierarchy of cells in tissue patches (Hinton, 1988). For both methods, the encoder f is parameterized using ResNet-18 and ViT-Tiny/16. ViTs receive 224×224 image tiles further divided into non-overlapping patches, processed through multi-head self-attention layers to learn tissue organization patterns. An online linear classifier is trained during SSL to provide pseudo-label feedback, enabling consistent assessment of the learned embeddings' utility for the downstream task.

3.4. Data Processing and Downstream Classification

Augmentations. Augmentation policies for SimCLR and DINO are presented are maintained from the original implementations. These augmentations, such as random resized crops, color jitter, and Gaussian blur, are tailored to enhance model robustness to various data augmentations, facilitating better generalization.

Occupancy Thresholds. Occupancy thresholding is employed to efficiently process large WSIs by focusing on tissue-containing regions, reducing computational load, and enhancing accuracy. For PANDA, an occupancy threshold $\Gamma = 0.5$ is used, while for TCGA-CRCk, $\Gamma = 0$. This thresholding ensures that only relevant tissue areas are analyzed, optimizing resource use and improving model performance.

Image Normalization. Normalization adjusts pixel values to align with the model's pretraining data distribution, improving stability and convergence during training. Although ImageNet normalization is standard for natural images, we compute dataset-specific statistics for histopathology images to minimize domain shift and improve model adaptability to medical imaging data.

Deep MIL for Classification. Deep MIL, used in our end-to-end workflow, facilitates slide-level classification by treating WSIs as bags of patches. Post-SSL pretraining, the model undergoes fine-tuning on TCGA-CRCk and PANDA datasets to adapt the learned features to specific slide-level tasks. The attention mechanism in MIL aggregates predictions from individual tiles, weighting them based on their relevance to the slide-level label.

GradCAM and Attention Rollout. To enhance interpretability, we employ visual XAI techniques, such as GradCAM for ResNets and attention rollout for ViTs. GradCAM generates class activation maps (CAMs) to highlight regions influencing model predictions, providing insights into the decision-making process. Attention rollout, applied to ViTs, quantifies attention flow across model layers, revealing which parts of the image influence the final prediction. These techniques ensure the model's transparency and facilitate clinical adoption by aligning model focus with pathologist insights.

4. Experiments and Results

For detailed technical specifications and additional training information related to our experimental setup, please refer to Appendix A. Additionally, refer to Appendix C for a fascinating perspective on our ImageNet pretrained baseline model![1]

4.1. Performance on TCGA-CRCk

Metric \ Experiment	SimCLR RN18	SimCLR ViT-T/16	DINO RN18	DINO ViT-T/16	LN ViT-T/16
Accuracy	**0.774 ± 0.032**	0.764 ± 0.050	0.544 ± 0.259	0.452 ± 0.235	0.592 ± 0.012
AUC	**0.805 ± 0.039**	0.774 ± 0.058	0.569 ± 0.092	0.485 ± 0.037	0.606 ± 0.034
F1	**0.588 ± 0.067**	0.564 ± 0.094	0.258 ± 0.236	0.248 ± 0.202	0.360 ± 0.023

Table 1: Performance on TCGA-CRCk using SSL pretrained ViTs and ResNets.

Metric \ Experiment	DINO ViT-B/16	DINO RN101	SimCLR ViT-B/16	SimCLR RN101
Accuracy	**0.604 ± 0.055**	0.573 ± 0.068	0.514 ± 0.077	0.473 ± 0.088
AUC	**0.610 ± 0.030**	0.588 ± 0.052	0.522 ± 0.048	0.497 ± 0.066
F1	**0.405 ± 0.063**	0.401 ± 0.086	0.336 ± 0.084	0.312 ± 0.096

Table 2: Performance on TCGA-CRCk using larger architectures.

In reference to Table 1, it is evident that the SimCLR ResNet18 model outperforms the others across all metrics, including accuracy, AUC, and F1 score. Furthermore, this model exhibits the lowest standard deviation, indicating a consistent performance. When comparing the performance of ResNet-18 and ViT-Tiny/16 models under SimCLR and DINO regimes, it is clear that SimCLR consistently outperforms DINO across all metrics. Moreover, the standard deviations of the DINO models, particularly for accuracy and F1 score, are significantly

1. Recommended after perusing all of Section 4.

higher, with the F1-score standard deviation nearly matching the mean. This suggests that the combination of DINO with ResNet-18 and ViT-Tiny/16 does not yield optimal results. Interestingly, the baseline model (ImageNet ViT-Tiny/16) is surpassed by both SimCLR models, while both DINO models underperform in comparison to the baseline.

In Table 2, it is noteworthy that the DINO models with larger architectures, namely ViT-Base/16 and ResNet-101, demonstrate improved performance compared to their counterparts with smaller architectures (as shown in Table 1). This suggests that larger architectures can significantly enhance the performance of models trained under the DINO regime. In particular, the DINO model with ViT-Base/16 outperforms the baseline model (highlighted in red in Table 1). However, it still falls short of the performance exhibited by the SimCLR model with ResNet18 (as per Table 1). This underscores the effectiveness of the SimCLR regime, particularly when combined with the ResNet-18 architecture[2].

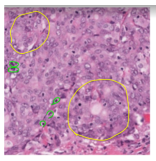

Figure 2: DINO RN18

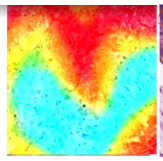

Figure 3: DINO ViT-T16

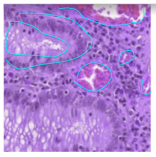

Figure 4: SimCLR RN18

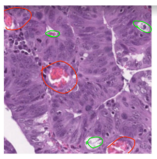

Figure 5: SimCLR ViT-T16

Ground-truth vs. CAM Interpretations.

Figure 2: The DINO ResNet-18 model fails to focus on the ROI needed for diagnosis, leading to misclassification. The ground truth shows compact acini with minimal lumen space, small cells with prominent nucleoli, and loss of basilar cells, all indicating prostate cancer ($y = 1, \hat{y} = 0$, *very inaccurate* CAM).

Figure 3: The ground truth image shows possible acinar cells or glandular structures with enlarged nucleoli and elongated nuclei, indicative of prostate adenocarcinoma. The DINO ViT-Tiny/16 model correctly identifies these high-density eosinophilic regions, matching the ground truth ($y = 1, \hat{y} = 1$, *very accurate* CAM).

Figure 4: The image displays cribriform areas with densely eosinophilic nuclei and cherry red cells, consistent with adenocarcinoma. While the SimCLR ResNet-18 model correctly identifies the left glandular region, it misses other crucial areas. The image contains multiple malignant indicators, complicating ROI selection ($y = 1, \hat{y} = 1$, *somewhat accurate* CAM).

Figure 5: The SimCLR ViT-Tiny/16 model identifies most ROIs, including cherry red nucleoli, and compact cells characteristic of adenocarcinoma. The only discrepancy is crystalloid structures in the lumen, circled in the ground truth image ($y = 1, \hat{y} = 1$, *accurate* CAM).

4.2. Performance on PANDA

In Table 3 The SimCLR models consistently outperform the DINO models in terms of Accuracy and AUC for both ResNet-18 and ViT-Tiny/16 architectures. Specifically, the ViT-Tiny/16 model trained with SimCLR exhibits superior performance across these metrics and also surpasses the baseline model.

Metric \ Experiment	SimCLR RN18	SimCLR ViT-T/16	DINO RN18	DINO ViT-T/16	I.N ViT-T/16
Accuracy	0.607 ± 0.026	**0.608 ± 0.008**	0.296 ± 0.010	0.563 ± 0.005	0.595 ± 0.015
AUC	0.880 ± 0.003	**0.881 ± 0.002**	0.638 ± 0.004	0.845 ± 0.002	0.867 ± 0.002
ISUP 0	**0.927 ± 0.014**	0.894 ± 0.014	0.793 ± 0.254	0.877 ± 0.009	0.885 ± 0.015
ISUP 1	0.580 ± 0.032	**0.605 ± 0.023**	0.166 ± 0.370	0.537 ± 0.015	0.543 ± 0.008
ISUP 2	**0.487 ± 0.033**	0.448 ± 0.049	0.161 ± 0.248	0.413 ± 0.016	0.405 ± 0.010
ISUP 3	**0.317 ± 0.050**	0.313 ± 0.019	0.000 ± 0.000	0.252 ± 0.023	0.272 ± 0.031
ISUP 4	0.402 ± 0.025	**0.447 ± 0.025**	0.111 ± 0.211	0.383 ± 0.042	0.353 ± 0.042
ISUP 5	**0.615 ± 0.021**	0.593 ± 0.004	0.000 ± 0.000	0.547 ± 0.015	0.600 ± 0.030
Binary F1	0.815	**0.856**	N/A	0.810	0.807

Table 3: Performance on PANDA using SSL pretrained ViTs and ResNets.

Metric \ Experiment	DINO ViT-B/16	DINO RN101	SimCLR ViT-B/16	SimCLR RN101
Accuracy	**0.637 ± 0.010**	0.319 ± 0.017	0.491 ± 0.033	0.264 ± 0.041
AUC	**0.892 ± 0.004**	0.678 ± 0.001	0.781 ± 0.017	0.613 ± 0.009
ISUP 0	**0.913 ± 0.016**	0.761 ± 0.087	0.772 ± 0.042	0.663 ± 0.072
ISUP 1	**0.605 ± 0.035**	0.072 ± 0.099	0.467 ± 0.060	0.027 ± 0.150
ISUP 2	**0.505 ± 0.075**	0.003 ± 0.002	0.363 ± 0.086	0.000 ± 0.004
ISUP 3	**0.318 ± 0.038**	0.150 ± 0.335	0.244 ± 0.059	0.063 ± 0.193
ISUP 4	**0.486 ± 0.023**	0.291 ± 0.399	0.382 ± 0.041	0.214 ± 0.330
ISUP 5	**0.620 ± 0.015**	0.311 ± 0.426	0.475 ± 0.039	0.252 ± 0.423
Binary F1	**0.873**	0.610	0.546	0.508

Table 4: Performance on PANDA using larger architectures.

The ISUP 0 class is classified with the highest accuracy compared to the other classes, with a significant margin. For the ISUP classes 1-5, which represent different stages of cancer, class 5 tends to be classified most accurately across all models, with the exception of DINO with ResNet-18. This observation suggests that the model might be confusing different stages of cancer with each other. To investigate

2. For further granular analysis over a sampled subset, please refer to Appendix B.

this, we computed the Binary F1 score by consolidating all ISUP classes 1-5 into a single class[3]. The resultant score was significantly higher than when the classes were treated separately, supporting the aforementioned hypothesis. For the DINO ResNet-18 model, it was not possible to compute a Binary F1 score as the scores for ISUP classes 3 and 5 were both zero, which would result in a division by zero during the computation. In Table 4, when DINO is paired with larger architectures, it exhibits improved performance compared to its smaller counterparts (ViT-Base/16 vs. ViT-Tiny/16 and ResNet-101 vs. ResNet-18). Notably, the accuracy of the DINO model with ViT-Base/16 is so high that it outperforms all other models in Table 3, and this trend is observed across most of the other metrics as well[4].

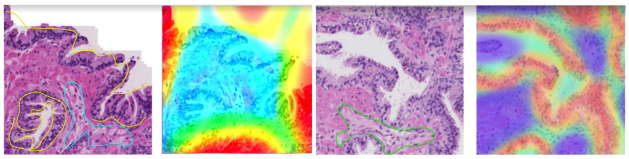

Figure 6: DINO RN18

Figure 7: DINO ViT-T16

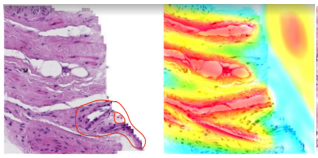

Figure 8: SimCLR RN18

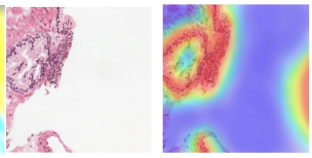

Figure 9: SimCLR ViT

Ground-truth vs. CAM Interpretations.

Figure 6: The acini structure appears normal, but eosinophilic uptake and uniform-like cell clumping suggest adenocarcinoma. Blue-marked areas in the ground truth image may indicate benign hyperplasia with basal cells, suggesting a less severe grade. The DINO ResNet-18 model fails to detect these ROIs, leading to a correct prediction for incorrect reasons ($y = 0, \hat{y} = 0$, *very inaccurate* CAM).

Figure 7: Loss of basal cells and irregular acini patterns suggest adenocarcinoma. The DINO ViT-Tiny/16 model highlights a uniform circular cell pattern but aligns more with benign hyperplasia. A green-marked area in the ground truth could indicate an invasive process, reflected by the cherry red nuclei. The model correctly classifies the image ($y = 0, \hat{y} = 0$, *somewhat accurate* CAM).

Figure 8: Large luminal areas suggest cystic hyperplasia. Irregular glandular cells and a red-circled area could signal malignancy due to decreased luminal space and eosinophilic nuclei. The SimCLR ResNet-18 model selects normal lumen spaces, missing critical ROIs. Despite this, the classification is correct but for the wrong reasons ($y = 0, \hat{y} = 0$, *somewhat inaccurate* CAM).

Figure 9: Irregular acini indicate benign tissue, and the red-marked area by the SimCLR ViT-Tiny/16 model may represent papillary structures. Lack of context or higher magnification limits meaningful conclusions. The wrong classification is likely due to insufficient detail, though the model identifies interesting regions ($y = 0, \hat{y} = 1$, *somewhat inaccurate* CAM).

4.3. Attention Weights Plot Demo

It is also crucial to acknowledge that the use of attention-based deep MIL enables us to identify specific patches within a bag that the model deems as ROIs within the broader WSI landscape. While GradCAM and attention rollout offer local granularity per patch, the attention weights plots below show us information about the patches most (and least) instrumental for a prediction made by the model.

The demo presents a correct positive ISUP prediction on the PANDA dataset. Specifically, Figure 10 displays the top 12 tiles that received the highest attention weights (left), while exhibiting the bottom 12 tiles with the lowest attention weights (right). It is interesting to note that the tiles with the lowest attention scores have a low occupancy rate, while a higher occupancy garners more attention from the ViT-Tiny/16 model. This degree of transparency, in conjunction with GradCAM and attention rollout, provides a comprehensive and fully interpretable solution for slide-level classification.

5. Discussion

The primary objective of this study was to perform a comparative analysis between ViTs and ResNets for slide-level classification in histopathology image analysis. Additionally, we investigated the impact of different SSL strategies—SimCLR and DINO—on model performance. Our key contributions are as follows:

1. SimCLR consistently performed well across smaller architectures for both TCGA-CRCk and

3. We used only the best cross validation run.
4. For further granular analysis over a sampled subset, please refer to Appendix B.

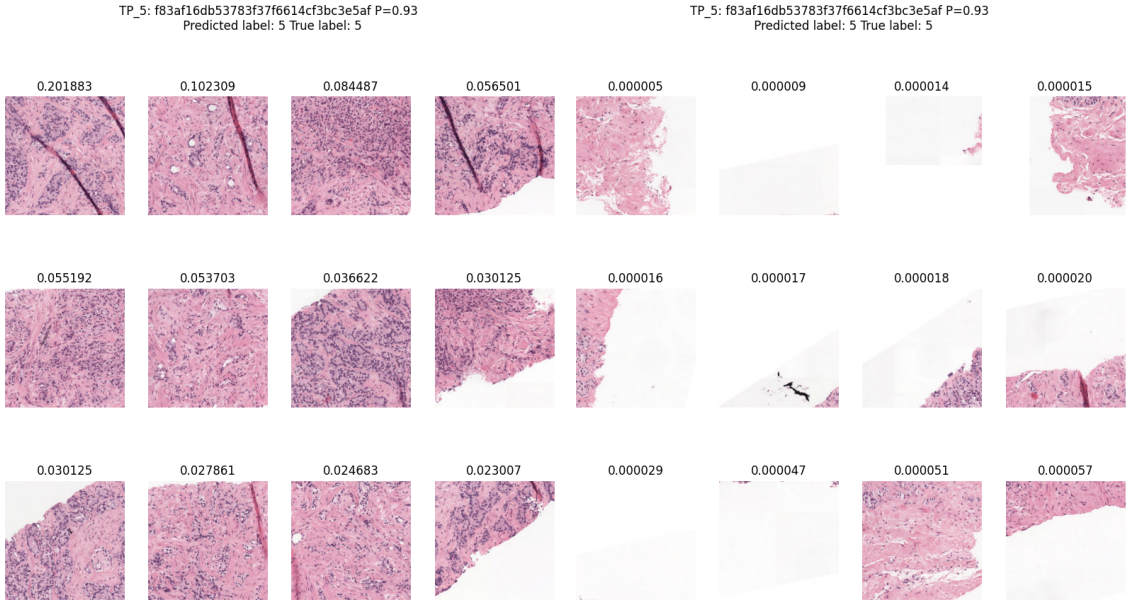

Figure 10: Deep MIL attention scores from SimCLR ViT-T/16 on a PANDA sample. Top 12 highest scores (left), and Top 12 lowest scores (right). The scores display the most relevant (or irrelevant) regions for prediction giving clinicians added interpretability within a clinical decision support system.

PANDA datasets, while DINO showed improvements with larger architectures. The reason DINO works better in the larger model regime is due to its self-attention mechanism, which better leverages the capacity of bigger architectures to capture long-range dependencies in the data.

2. ResNets did not perform optimally in the DINO regime, regardless of architecture size, highlighting a potential limitation in the compatibility of ResNets with DINO for histopathology tasks. ResNets may struggle with DINO due to architectural differences in feature extraction. DINO relies heavily on attention mechanisms, which may align better with ViTs' self-attention layers than with ResNets' convolutional layers. Furthermore, larger ViTs improved overall performance.

3. ViTs outperformed ResNets on the PANDA dataset but struggled on TCGA-CRCk until architecture size increased, without surpassing the smaller ResNet-18 in the SimCLR regime; while SimCLR dropped in performance in tandem with larger architectures.

4. Our interpretability analysis using GradCAM and attention rollout revealed that ViTs offer better visualization of histopathology features, making them a promising tool for diagnostic applications. The attention weights plot demo using Deep MIL (Figure 10) is another interpretability tool offered for ROI analysis.

The natural trade-offs between ResNets and ViTs, and between SimCLR and DINO, can inform model selection based on clinical constraints like computational resources and the need for transparency and granularity in decision-making. Different encoders naturally have distinct impacts in contrastive and self-distillation SSL methods like SimCLR and DINO. ResNet, a CNN-based encoder, and ViT, a transformer-based model, each learn visual features in unique ways—ResNet through local receptive fields and hierarchical feature extraction, and ViT through self-attention and global context modeling across image patches. This gives rise to what we call the *self-supervision regime and encoder fit* (SSL-encoder fit) as our study opens several avenues for future research:

SSL-encoder Fit. Our results suggest a strong affinity between ViTs and DINO, as well as between ResNets and SimCLR. This observation raises an in-

triguing question regarding the existence of a regime-agnostic SSL approach for histopathology tasks. Further research is necessary to explore this concept of "SSL-encoder fit" and its implications for model performance across different architectures. Interestingly, this phenomenon is rooted in information theory and calls for a more comprehensive, large-scale, and holistic exploration, making it an extremely exciting avenue for further study Shwartz Ziv and LeCun (2024).

Scale Optimization. Further experiments with larger architectures and alternative SSL methods (e.g., BYOL, MAE) are necessary to validate whether ViTs consistently outperform ResNets in histopathology image analysis (Grill et al., 2020; He et al., 2021). Exploring newer architectures like Swin Transformers or ConvNext could provide additional insights into model performance (Liu et al., 2021, 2022).

Data Optimization. The quality of the datasets, particularly the inclusion of non-informative patches (e.g., white space in TCGA-CRCk), likely impacted model performance. The lack of an occupancy threshold parameter for optimizing TCGA-CRCk may have reduced the effectiveness of DINO's local-to-global correspondence mechanisms, contributing to its suboptimal performance with smaller architectures. Therefore, curating diverse histopathology datasets with optimized parameters (e.g., occupancy thresholds, magnification levels) will help create robust benchmarks. Moreover, patch-level classification should be explored to determine whether ViTs or ResNets dominate across multiple levels of granularity in histopathology tasks.

Recommendations for Potential Adoption. SimCLR with ResNet models consistently performs well, especially with smaller architectures like ResNet-18. This combination could be optimal for institutions working with limited computational resources or smaller datasets, where the simpler, more efficient architecture can still deliver strong performance. We recommend SimCLR with ResNet as a viable approach when batch sizes are carefully controlled, and negative samples can be sufficiently provided. We observed that DINO's performance improves significantly with larger ViT architectures (e.g., ViT-B/16), making it more suitable for large-scale, complex histopathology tasks such as multi-class cancer grading. This setup is recommended for larger datasets and more computationally demanding tasks, where DINO's attention mechanisms allow for better feature extraction and improved interpretability. The key practical advantages of using ViTs with DINO, as shown by our experiments, is the enhanced interpretability of the results through attention mechanisms. This could be particularly valuable for pathologists seeking insights into which regions of the histopathological image are most relevant to the model's decision. Additionally, the MIL weighted attention plots demo (Figure 10) provides a broader view of how the model prioritizes different patches, further assisting in understanding model focus at a slide-level, making it an even more effective tool for clinical validation and trust for pathologists.

6. Conclusion

In conclusion, our main goal was to systematically evaluate and compare two major SSL strategies (SimCLR for contrastive learning and DINO for self-distillation) across two commonly used encoders (ViTs and ResNets) for histopathology image analysis. While ViTs demonstrate promise in histopathology image analysis, particularly in the DINO regime, further experimentation and exploration are required to definitively conclude whether ViTs can consistently outperform ResNets across all self-supervised learning scenarios. In summary, our findings hold practical value for clinical teams engaged in translational work in histopathology, providing guidance to streamline their search for effective methods tailored to their specific tasks.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

Mahul B. Amin, Frederick L. Greene, Stephen B. Edge, Carolyn C. Compton, Jeffrey E. Gershenwald, Robert K. Brookland, Laura Meyer, Donna M. Gress, David R. Byrd, and David P. Winchester. The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: A Cancer Journal for Clinicians*, 67(2):93–99, January 2017. doi: 10.3322/caac.21388. URL https://doi.org/10.3322/caac.21388.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, July 2019. doi: 10.1038/s41591-019-0508-1. URL https://doi.org/10.1038/s41591-019-0508-1.

Grant D Carlson, Christina B Calvanese, Hillel Kahane, and Jonathan I Epstein. Accuracy of biopsy gleason scores from a large uropathology laboratory: Use of a diagnostic protocol to minimize observer variability. *Urology*, 51(4):525–529, April 1998. doi: 10.1016/s0090-4295(98)00002-8. URL https://doi.org/10.1016/s0090-4295(98)00002-8.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

The PANDA challenge consortium. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 2022. doi: 10.1038/s41591-021-01620-2. URL https://doi.org/10.1038/s41591-021-01620-2.

Richard J. Chen and Rahul G. Krishnan. Self-supervised vision transformers learn visual concepts in histopathology, 2022.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Geoffrey E Hinton. Representing part-whole hierarchies in connectionist networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 48–54, 1988.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018.

Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In Steven C. H. Hoi and Wray Buntine, editors,

Proceedings of the Asian Conference on Machine Learning, volume 25 of *Proceedings of Machine Learning Research*, pages 253–268, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/liu12b.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images, 2020.

A G Nicholson, L J Perry, P M Cury, P Jackson, C M McCormick, B Corrin, and A U Wells. Reproducibility of the WHO/IASLC grading system for pre-invasive squamous lesions of the bronchus: a study of inter-observer and intra-observer variation. *Histopathology*, 38(3): 202–208, March 2001. doi: 10.1046/j.1365-2559.2001.01078.x. URL https://doi.org/10.1046/j.1365-2559.2001.01078.x.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017.

Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zieliński. Kernel self-attention in deep multiple instance learning, 2021.

Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from he whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79, jul 2022. doi: 10.1016/j.media.2022.102464. URL https://arxiv.org/abs/2107.09405.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007%2Fs11263-019-01228-7.

Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3), 2024. ISSN 1099-4300. doi: 10.3390/e26030252. URL https://www.mdpi.com/1099-4300/26/3/252.

Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics*, 9(1):38, January 2018. doi: 10.4103/jpi.jpi_53_18. URL https://doi.org/10.4103/jpi.jpi_53_18.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, feb 2018. doi: 10.1016/j.patcog.2017.08.026. URL https://doi.org/10.1016%2Fj.patcog.2017.08.026.

Conghao Xiong, Hao Chen, Joseph Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2018.

Appendix A. Specifications

Models and Hyperparameters. The set of models used for pretraining, fine-tuning, and downstream classification can be found below. All pertinent parameters and hyperparameters have been documented. It is important to note that during the pretraining phase, all parameters were maintained the same as the vanilla parameters used in SimCLR and DINO. By keeping these parameters constant, we could effectively assess the impact of other modifications and strategies on the performance of the models during the pretraining and fine-tuning stages. The relevant models and their parameters for the fine-tuning phase and downstream task are seen in Table 5 and Table 6.

Model	Trainable Parameters	Non-Trainable Parameters
DINO ViT-T/16 (PANDA)	26K	5.5M
DINO RN18 (PANDA)	68.9K	11.2M
DINO ViT-T/16 (PANDA)	25K	5.5M
DINO RN18 (CRCk)	66.3K	11.2M
SimCLR ViT-T/16 (PANDA)	26K	5.5M
SimCLR RN18 (PANDA)	68.9K	11.2M
SimCLR ViT-T/16 (CRCk)	25K	5.5M
SimCLR RN18 (CRCk)	26K	11.2M

Table 5: Total number of parameters for all models

Parameter	Value
cross_val_count	5
batch_size	32
max_epochs	100
max_bag_size	1000
l_rate	0.0005
optimizer_type	Adam
num_transformer_pool_heads	4
num_transformer_pool_layers	4
pool_hidden_dim	128
pool_type	AttentionLayer (Ilse et al., 2018)
momentum	0.6
weight_decay	0.0001
tile_size	224

Table 6: Fine-tuning/downstream phase: important parameters for all models.

Validation Images. During the DINO pretraining process, we adopted a validation strategy where the same data transformations were applied to both the training and validation images. The objective was to observe if this approach would lead to smoother AUC and loss curves, similar to what is generally observed with SimCLR, which also validated on augmented images similar to its training data. Surprisingly, this strategy resulted in an average increase of 0.06 in the AUC of DINO. As a result, we decided to fix this validation strategy for the final pretraining of models, as it seemed to improve the performance of DINO during validation. Note that training and validating on images from the same distribution was favorable likely because of the model updating its weights based on images it would recongnize in the validation step that were similar to the training data. The idea is to pretrain on low-level features found in the augmentations and later potentially fine-tune on the high-level features that it may find while training on unaugmented patches downstream.

Appendix B. A More Granular Look

B.1. TCGA-CRCk Confusion Matrices

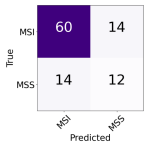

Figure 11: DINO RN18

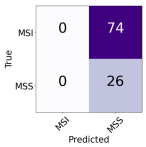

Figure 12: DINO ViT-T/16

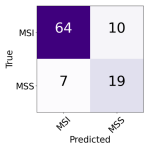

Figure 13: SimCLR RN18

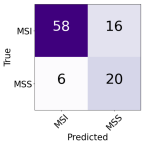

Figure 14: SimCLR ViT-T/16

In reference to the confusion matrices in Figures 11, 12, 13, and 14, a consistent trend emerges that aligns with the findings in Section 4.3.1. Both SimCLR models (Figures 13 and 14) exhibit a higher number of true positives and true negatives, indicating a more balanced model performance compared to the DINO models (Figures 11 and 12). The DINO models appear to overfit on the MSS class, particularly evident in the DINO model with ViT-Tiny/16 (Figure 12), where each patch is classified as MSS. The most balanced and best-performing model is achieved with SimCLR using ResNet18, as it yields the highest number of true negatives and true positives.

MSS cancers are generally more challenging to detect and treat, while MSI cancers are easier to identify and manage. Given this, it is crucial that true MSS cases are accurately classified. If an easier-to-treat cancer (MSI) is misclassified as a harder-to-treat cancer (MSS), the consequences are less severe than

if a harder-to-treat cancer (MSS) is misclassified as an easier-to-treat cancer (MSI). The latter scenario could potentially delay the diagnosis and treatment process, which could be detrimental given the already challenging nature of treating MSS cancers.

From this perspective, the SimCLR models (Figures 13 and 14) still outperform the DINO models, even though the DINO model with ViT-Tiny/16 (Figure 12) has the lowest false MSI rate. However, this is not a proper classification as all instances are classified into the same class. While it was previously suggested that the SimCLR model with ResNet18 performed the best, considering this new perspective, the SimCLR model with ViT-Tiny/16 might be superior. This model has a higher false MSS rate, but a lower false MSI rate, as can be seen in Figures 11 and 12 when comparing the two. Note that the best AUC-based cross validation runs for each model were used for comparison amid their respective confusion matrices.

B.2. PANDA Confusion Matrices

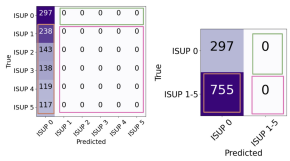

Figure 15: DINO RN18

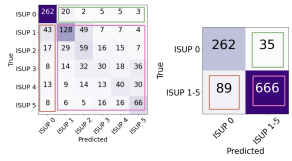

Figure 16: DINO VIT-T16

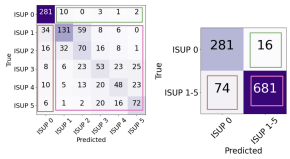

Figure 17: SimCLR RN18

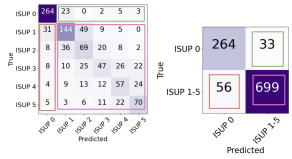

Figure 18: SimCLR ViT-T16

From Figure 15, DINO ResNet-18 exhibits a similar issue to the DINO ViT-Tiny/16 for the CRCK dataset, as it classifies all images into a single class. These multiclass matrices reveal that the models often misclassify different levels of ISUP for each other, confirming our previous suspicion. Hence, we have binarized the multiclass confusion matrices for a simpler analysis. In the confusion matrices in Figures 16, 17, and 18, they all display a similar pattern, where they predict classes ISUP 0 and ISUP 1-5 with roughly equal accuracy. However, each model may exhibit a slight bias towards one of the classes. The choice of the best model would then depend on the specific priorities. Given that ISUP 0 represents the absence of cancer, it can be argued that correctly predicting ISUP 1-5 should be prioritized, as misclassifying a cancer patient as cancer-free could have severe consequences. Therefore, when considering the confusion matrices, the SimCLR ViT-Tiny/16 model (Figure 18) would be the best choice, as it has the lowest false ISUP 0 rate.

Appendix C. A Pespective on ImageNet Pretraining

A likely reason why a ViT-T/16 model pretrained on ImageNet outperforms DINO pretraining (with ViT-T/16 or RN18) in the small architecture regime is due to the nature of the supervision used during pretraining. ImageNet pretraining is based on fully supervised learning, where the model is explicitly trained to classify images into thousands of well-defined categories. This type of training provides strong and highly specific signal, especially when dealing with relatively small models like ViT-T/16 or RN18, which have limited capacity to learn complex representations. In contrast, DINO relies on self-supervised learning, where the model must learn to extract useful features without direct supervision. While DINO is powerful with larger architectures that have sufficient capacity to handle its self-distillation and attention mechanisms, smaller architectures like ViT-T/16 and RN18 may struggle to fully capture meaningful patterns from unlabelled data. Without explicit labels, these smaller models might fail to learn as rich a feature space as the ImageNet-supervised model, which benefits from being exposed to a diverse and labeled dataset. Thus, the combination of limited model capacity in the small architecture regime and the more challenging nature of self-supervised pretraining likely explains why ImageNet-supervised ViT-T/16 outperforms DINO-pretrained models in this context.