Beyond Labels: A Self-Supervised Framework with Masked Autoencoders and Random Cropping for Breast Cancer Sub-type Classification

Marco Dossena[1], **Christopher Irwin**[1], **Annalisa Chiocchetti**[2], **Luigi Portinale**[1]
[1]Computer Science Institute, DiSIT, University of Piemonte Orientale, Alessandria (Italy)
[2]Department of Health Sciences, University of Piemonte Orientale, Novara (Italy)
[1]Computer Science Institute, DiSIT, University of Piemonte Orientale, Alessandria (Italy)
marco.dossena@uniupo.it, christopher.irwin@uniupo.it, annalisa.chiocchetti@uniupo.it, luigi.portinale@uniupo.it

Abstract

This work addresses the problem of breast cancer sub-type classification using histopathological image analysis. We utilize masked autoencoders (MAEs) based on Vision Transformer (ViT) to learn, through Self-Supervised Learning, embeddings tailored to computer vision tasks in this domain. Such embeddings capture informative representations of histopathological data, facilitating feature learning without extensive labeled datasets. During pre-training, we investigate employing a random crop technique to generate a large dataset from whole-slide images automatically. Additionally, we assess the performance of linear probes for multi-class classification tasks of cancer sub-types using the representations learned by the MAE. Our approach aims to achieve strong performance on downstream classification task, by leveraging the complementary strengths of ViTs and autoencoders. We evaluate our model's performance on the BRACS and BACH datasets and compare it with existing benchmarks.

Introduction

Histopathological image analysis plays a critical role in clinical applications such as cancer diagnosis. Whole-slide images (WSIs) offer high-resolution views of entire tissue sections, enabling comprehensive evaluation by pathologists. However, manual analysis of WSIs is time-consuming and prone to inter-observer variability. Deep learning models have emerged as powerful tools to automate histopathological image analysis, offering the potential for faster, more consistent, and potentially more accurate diagnoses (Van der Laak, Litjens, and Ciompi 2021). Given the high-resolution nature of WSIs, it is interesting to adopt a localized analysis approach. This involves the extraction of image patches, in order to address tasks like classification at the level of bag of tissue regions; to this end, a multiple instance learning framework could be naturally applied (Wang et al. 2024). Furthermore, the reduced size of the extracted patches facilitate the application of deep learning models for computer vision tasks, including Convolutional Neural Networks (CNN) (LeCun et al. 1989) and Vision Transformer (ViT) architectures (Dosovitskiy et al. 2021), for various objectives such as tissue classification and cell segmentation.

CNNs have become the most used approach in this domain due to their ability to capture spatial relationships within images (Srinidhi, Ciga, and Martel 2021; Hou et al. 2016). Architectures such as VGG, ResNet, and Inception excel at learning hierarchical features directly from raw image data, making them ideal for tasks like tissue classification, tumor segmentation, and cell detection. However, CNNs struggle to capture long-range dependencies (i.e., overall features) within complex tissues, especially when the networks depth increases (Qiong et al. 2025).

A promising alternative is offered by Graph Neural Networks (GNN). GNNs represent tissue structures as graphs (Zhou et al. 2019; Aygüneş et al. 2020), where nodes represent cells and edges depict their relationships. This allows GNNs to effectively model complex interactions between cells, making them particularly useful for analyzing cell-to-cell communication or studying the spatial distribution of different cell types. Alternatively, for tasks aiming to classify multiple bag of regions under a single label, the graph representation can be constructed at the patch level. The downside of this kind of approach is the heavy preprocessing step needed to build the required graph structure (Pati et al. 2022).

ViTs represent another interesting approach. Unlike CNNs, ViTs process image patches directly, leveraging transformer techniques to learn global dependencies across the entire image (Gul et al. 2022; Wang et al. 2021a). This approach proves advantageous for tasks requiring the analysis of intricate tissue patterns, and overcomes the limitations imposed by pre-defined filter sizes in CNNs.

However, independently of the adopted learning architecture, because of the lack of large-scale annotated datasets, the field of computer-aided medical imaging has witnessed a widespread adoption of transfer learning especially from ImageNet. As a matter of fact, histological images exhibit complex and specific features, related to cellular structures, tissue morphology and staining patterns, which may not be suitably captured and dealt with by models pre-trained on a very general dataset such as ImageNet (Filiot et al. 2023).

In this paper, we aim at exploiting encoding/decoding features of ViT, in order to achieve a significant latent representation enabling an accurate classification of tissue regions or WSIs via Self-Supervised Learning (SSL), avoiding the need for large annotated datasets. We propose a reconstruc-

tion framework called *Histopathological Masked AutoEncoder* (HMAE) followed by a simple classifier. The core of the architecture is a ViT-based auto-encoder, where the construction of the latent space is achieved through the self-supervised objective of reconstructing the original image. By masking a significant portion of image patches during input, the encoder is forced to identify increasingly intricate patterns in the remaining data to reconstruct the complete image. Finally, a classification layer based on a simple MLP is applied after the training of the transformer, in order to output the predicted class. In this way, the ViT model is never exposed to the image class labels during the learning process. Figure 1 shows the proposed pipeline and architecture that will be detailed in the following.

Methodology and Related Works

This section discusses the synergy between ViTs and masking techniques for image reconstruction using a Masked Autoencoding framework (MAE) (He et al. 2022). We start by investigating the theoretical foundations of ViT and masked image encoding, exploring their architectural details and functionalities. Next, we describe the process of acquiring and pre-processing a dataset suitable for MAE training. This dataset will be extracted from whole slide images (WSI) obtained from a reference dataset containing histopathological images for patients under examination for breast cancer. Finally, we discuss the methodology employed to leverage the feature representations (embeddings) learned by the ViT model for the task of cancer and sub-cancer classification.

Vision Transformers

Vision Transformers (ViTs) (Dosovitskiy 2020) represent a paradigm shift in computer vision, achieving state-of-the-art results on image classification tasks while departing from the traditional dominance of CNNs. Unlike CNNs that rely on hand-crafted filters for feature extraction, ViTs leverage the transformer architecture, originally proposed in Natural Language Processing (NLP) (Vaswani 2017). The key idea is to split the input image into fixed-size patches. Such patches are then embedded into a vector representation and fed into an encoder block. A self-attention mechanism is then exploited, in order to learn long-range dependencies between different parts of the image, allowing the model to capture global context crucial for classification. A (masked) auto-encoder architecture can the be built using a ViT as a backbone.

Masked Encoder

The first step consists of dividing the image into equally sized non-overlapping patches. Then, instead of using the whole image, a given number of patches is randomly picked without replacement, while the rest is hidden (masking) (He et al. 2022). Patch selection is performed randomly across the whole image, in order to avoid favoring the image center (center bias).

As originally proposed in (He et al. 2022), we decided to keep unmasked only a small percentage of the original images, namely 25% of all the patches. This is done in such a way that it is hard for the model to guess what is missing by just looking at the nearby patches. Having very few unmasked patches allows us to design a more efficient system for processing the image, as we will explain next. Finally, the resulting masked image is used as input for a ViT encoder (see Figure 1).

Masked Decoder

After the encoding phase, we have two sets of information: encoded data for the visible patches and special "mask tokens" (Devlin 2018). Such mask tokens represent missing patches that the model needs to predict. A learned positional encoding is then added to both the encoded patches and mask tokens.

This information is passed through another series of transformer blocks, acting as a decoder. The decoder is only used during training to learn how to fill in the missing patches. It does this by predicting the actual pixel values for each masked area. The decoder's output is a vector of pixel values representing a patch, and the final step is to put all these patches back together to form a complete reconstructed image. Again, this whole process is shown in Figure 1. To measure the accuracy of the reconstruction process we compare the reconstructed image with the original one pixel by pixel by using the mean squared error (MSE).

Reference Datasets

Having decided the learning architecture, we have then considered some reference datasets for evaluating the classification capabilities in terms of histopathological images for breast cancer. The datasets we have considered in the present work are described in the following.

BRACS. The BReAst Carcinoma Subtyping (BRACS) dataset (Brancati et al. 2022) is a collection of digital images used to study breast lesions. It includes 547 WSIs, which are high-resolution scans of entire tissue samples. Additionally, 4539 smaller, more specific areas called regions of interest (RoIs) are extracted from these whole-slide images. Each WSI and its corresponding RoIs are carefully examined and labeled by three pathologists. These categories encompass three main lesion types, that is three different classes: *benign* (healthy tissue), *malignant* (cancerous tissue), and *atypical*. Further details are provided by seven subcategories within these main types, allowing for a more precise understanding of the specific lesion (see Figure 2).

BACH. The BreAst Cancer Histology (BACH) dataset (Aresta et al. 2019) is a significant resource for researchers developing computer algorithms to automatically diagnose breast cancer. It consists again of a collection of digitized WSIs from breast biopsies. The RoIs extracted from each WSI are labeled according to four different classes: *normal tissue*, *benign tumors*, *in situ carcinoma* (precancerous cells), and *invasive carcinoma* (cancerous cells). The dataset contains a total of 400 RoIs (100 RoIs for each class).

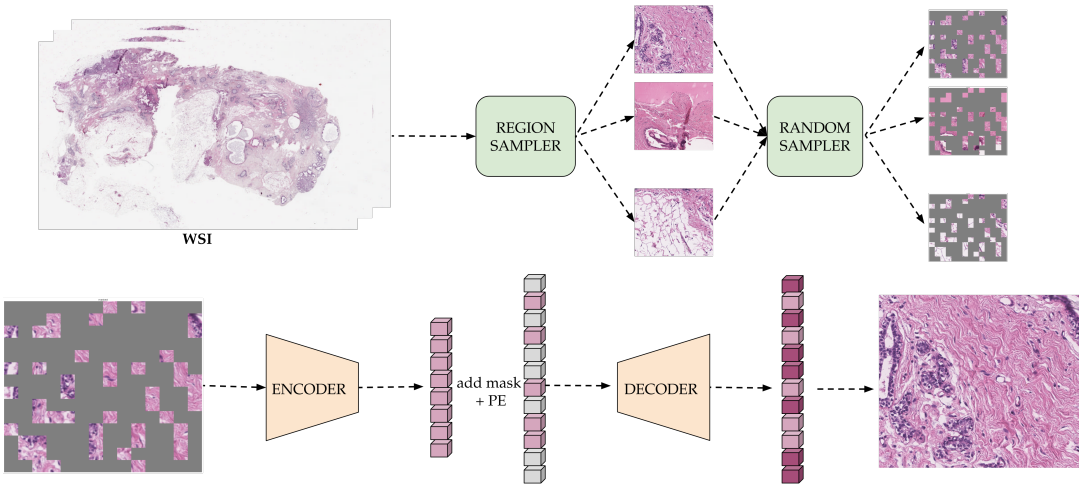

Figure 1: **HMAE architecture**. (top) Tissue regions are randomly sampled from the original image (WSI). Subsequently, a random mask is applied, occluding 75% of the image data. (down) The autoencoder architecture receives the masked image as input and aims to reconstruct the hidden regions.

Sampling Pipeline

The ultimate goal is to distinguish tumors from healthy tissues and to further classify tumor subt-ypes. This requires the model to learn informative representations of the tissues. To achieve this, we randomly extract a large number of unlabeled image regions from each WSI. This process aims to capture a diverse pool of tissue regions encompassing both tumor and non-tumor areas.
The steps for extracting an image from WSI are listed below:

Region Selection. A square-shaped image patch is chosen from the WSI. The side length of the region is determined by sampling from a normal distribution. The mean and standard deviation of this distribution is computed based on the size statistics of previously annotated RoIs in the BRACS dataset.

Region Quality Control. The average variance is computed for the pixel intensities within the extracted patch. Since this is a measure of dispersion relative to the mean, if the average variance is greater than a predefined threshold, the region is considered informative and included in the dataset. This step aims to exclude uninformative regions, such as borders and background areas, which often exhibit low variability with predominantly white or black pixels and are irrelevant for the task.

Annotated RoI Classification

This task investigates the ability of the proposed model to classify annotated RoIs within WSIs for tissue type classification. The approach is divided into two-steps. First, RoIs are fed into the model, which utilizes MAE to generate informative feature vectors for the patches. Subsequently, a mean aggregation of these patch-level embeddings is performed in order to obtain a single vector for every RoI. These embeddings capture the essential characteristics learned by the model from the data. Secondly, in order to carry out the classification task, we resort to an MLP with one hidden layer. The MLP utilizes the RoI embeddings as input features, where each class corresponds to a specific tissue type present within the RoIs.

In the following we will describe the experimental framework that has been set to evaluate the approach with the objective that, by successfully classifying different tissue types solely based on the embeddings, we can support the claim that the learned features effectively capture relevant diagnostic information within the RoIs.

The advantage is that RoI representations can be learned in a self-supervised manner without resorting to extensive labeling. Labels are only needed to train the classifier: the claim is that very few training examples are needed to perform the latter task when the embedding learned via SSL are good representative of the original RoIs.

Experimental Framework

A first evaluation has concerned the capability of the method previously described to correctly perform breast cancer subtype classification using histopathological hematoxylin and eosin (H&E) tissue images. To perform this task we have considered the BRACS dataset previously described. Following the BRACS group's categorization (see Figure 2), we have designed two multi-class classification experiments.

The first experiment differentiates cancerous from non-cancerous tissues, additionally incorporating an intermediate class for atypical cases, resulting in a 3-class classification problem. The second experiment extends the classification by attempting to categorize the tissues to a lower granularity level, namely into the distinct sub-types of the BRACS hierarchy (i.e., a 7-class classification problem).

Moreover, in order to test the generalization capabilities of the HMAE architecture we have considered the second reference dataset previously introduced, namely the BACH

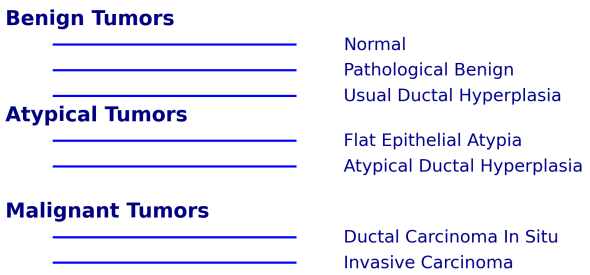

Figure 2: Class taxonomy of BRACS dataset.

dataset containing 400 breast cancer histopathology images classified into four categories. It is worth noting that the HMAE model was not trained on this dataset; instead, it was used in a frozen state to extract feature representations (embeddings) from the images. The extracted embeddings were then used to train an MLP classifier on the four classification labels specific to the BACH dataset.

Experimental setup. In all the evaluations, the corresponding dataset has been split into training, validation, and test sets using a 70/10/20 split. To account for potential variability, each classification experiment was run 100 times, and the average value for each chosen metric was reported. The MAE training was performed on a single Nvidia A40 GPU, taking a total time of 32 hours to complete.

Cancer Classification

The first evaluation concerned the task of classifying tissues from BRACS images as cancerous, non-cancerous or atypical. We have benchmarked HMAE with respect to several state-of-the-art approaches to this problem and the results are shown in Table 1, reporting the resulting AUC, and in Table 2, reporting the resulting weighted F1-score (for a 3-class classification task). For a correct comparison, all models employed a ViT-S/16 encoder architecture. This facilitated a comprehensive evaluation against various baseline models, leveraging the groundwork established in (Zhang et al. 2024).

These models use either an attention mechanism or a ViT to correlate different region of a tissue image. They then classify the tissue by considering the relationships between these regions (the first two in the table perform the maximum and the average between the image patches embedding respectively).

In both tables, the best performing approach is shown in bold and the HMAE results are background colored. Since HMAE does not perform as the best, we have tested whether there are significant differences in the measured performance scores with respect to the top performer approach (ACMIL). The p-values computed in a T-test at the 95% confidence level are reported in Table 3. Since p-values are greater than 0.05 we can conclude that, with a 95% confidence level, there is no significant difference between HMAE and ACMIL in terms of both WF1-score and AUC.

Model	AUC
Max-pooling	0.823±0.033
Mean-pooling	0.739±0.007
Clam-SB (Lu et al. 2021)	0.863±0.005
TransMIL (Shao et al. 2021)	0.841±0.006
DSMIL (Li, Li, and Eliceiri 2021)	0.816±0.028
DTFD-MIL (Zhang et al. 2022)	0.870±0.022
IBMIL (Lin et al. 2023)	0.871±0.014
MHIM-MIL (Tang et al. 2023)	0.865±0.017
ABMIL (Ilse, Tomczak, and Welling 2018)	0.866±0.029
ACMIL (Zhang et al. 2024)	**0.888±0.010**
HMAE	0.866±0.003

Table 1: Cancer Classification (3 classes): AUC

Model	WF1-score
Max-pooling	0.596±0.029
Mean-pooling	0.522±0.038
Clam-SB (Lu et al. 2021)	0.631±0.034
TransMIL (Shao et al. 2021)	0.631±0.030
DSMIL (Li, Li, and Eliceiri 2021)	0.577±0.028
DTFD-MIL (Zhang et al. 2022)	0.612±0.080
IBMIL (Lin et al. 2023)	0.645±0.041
MHIM-MIL (Tang et al. 2023)	0.625±0.060
ABMIL (Ilse, Tomczak, and Welling 2018)	0.680±0.051
ACMIL (Zhang et al. 2024)	**0.722±0.030**
HMAE	0.704±0.009

Table 2: Cancer Classification (3 classes): weighted F1-score

Sub-type Cancer Classification

A second evaluation has concerned the sub-type cancer classification in the BRACS dataset. This task represents an even greater challenge than previous efforts due to the often subtle morphological distinctions between some cancer types. Furthermore, the inherent heterogeneity of the training data, which includes a significant percentage of non-tumor tissue, introduces an inherent class imbalance within the latent representation space. The performance of our model was evaluated against the benchmark reported in (Stegmüller et al. 2023) under identical experimental conditions (7-class classification task, weighted F1-score for evaluation).

Most of the models used for this comparison differ from those employed in cancer classification (reported on Table 1 and Tables 2) since benchmarks for cancer and sub-type cancer classification are different; in fact, models for cancer classification (discussed in (Zhang et al. 2024)) would require training from scratch specifically for sub-type classification, whereas our model does not. Results are presented in Table 4 (best performance in bold and HMAE results background colored).

In this case, statistical significance analysis between HMAE and the top-2 performer approaches reveals a significant difference (at 95% confidence level) in the computed WF1-scores (see Table 5). In order to get a deeper understanding of such differences, we investigated the behavior, among the top-3 performers (ScoreNet, HACT-Net and HMAE), in terms of F1-score on each label (i.e., nor-

	ACMIL	HMAE	p-value
F1-score	0.722	0.704	0.2598
AUC	0.888	0.866	0.1728

Table 3: Cancer Classification (3 classes): statistical difference (95% confidence)

Model	WF1-score
CLAM-MB/B (Lu et al. 2021)	0.548±0.010
CGC-Net (Zhou et al. 2019)	0.436±0.005
Patch-GNN (Aygüneş et al. 2020)	0.521±0.006
TG-GNN (Pati et al. 2020)	0.559±0.001
CG-GNN (Pati et al. 2020)	0.566±0.013
HACT-Net (Pati et al. 2020)	0.615±0.009
TransPath (Wang et al. 2021b)	0.567±0.02
TransMIL (Shao et al. 2021)	0.575±0.007
ScoreNet (Stegmüller et al. 2023)	**0.644±0.009**
HMAE	0.578±0.015

Table 4: Sub-type cancer classification: weighted F1-score

mal tissue and sub-cancer categories). Results are reported on Table 6 (best results shown in bold as usual).

Previous exposure to a larger proportion of non-cancerous tissue during the training phase appears to have influenced the HMAE prediction distribution. This is reflected by the best F1-score achieved by HMAE for the "normal" class compared to both the benchmark models and other classes within this investigation. However, performance on the remaining classes, particularly the "Invasive" class with the highest F1-score, remains comparable to previously tested models.

In order to be more detailed about this aspect, we have performed a statistical significance test also in this case, and results are reported in Table 7. Bold values represent situations where HMAE performs either as significantly the best ("normal" class) or with no significant difference with respect to the other approaches. From these results we can conclude that, despite a potential bias towards non-cancerous tissue introduced by the training data, the model retains in general the ability to discriminate effectively between different cancer sub-types, quite often in a comparable way with respect to the best performing approaches.

Generalization Capabilities

To evaluate the representational capacity of our model, we have investigated its ability to generalize to data coming from completely unseen WSIs. This has been achieved by assessing its performance on a classification task using a dataset (BACH) not included in the training phase (tumor and non-tumor tissues, further categorized into four distinct

	ScoreNet	HACT-Net
HMAE	2.21E−13	3.58E−05

Table 5: Sub-cancer Classification (7 classes): p-values at 95% confidence

Label	ScoreNet	HACT-Net	HMAE
Normal	0.646±0.022	0.616±0.021	**0.683±0.022**
Benign	**0.540±0.022**	0.475±0.029	0.485±0.020
UDH	**0.484±0.022**	0.436±0.019	0.445±0.070
ADH	**0.474±0.024**	0.404±0.025	0.301±0.015
FEA	**0.779±0.007**	0.742±0.014	0.702±0.018
DCIS	0.629±0.020	**0.664±0.026**	0.633±0.019
Invasive	**0.910±0.014**	0.884±0.002	0.893±0.015

Table 6: Single class classification F1-score (based on top-3 models in Table 4).

Label	ScoreNet	HACT-Net
normal	**0.0198**	**0.0198**
benign	3.00E−04	**0.5779**
UDH	**0.2976**	**0.8078**
ADH	7.41E−32	6.00E−12
FEA	1.00E−14	6.00E−4
DCIS	**0.0594**	**0.0594**
Invasive	**0.1047**	**0.2447**

Table 7: Single-class performance: p-values between HMAE and top-2 performers

classes). In particular, the latent representation learning is performed using BRACS, then during inference each image from BACH is used as input to the HMAE architecture, the latent representation is produced and the classifier (now trained on a subset of BACH for a 4-class classification task) can finally predict the class. As in the previous experiments, the corresponding dataset has been split into training, validation, and test sets using an 70/10/20 split. Results (in terms of weighted F1-score) are shown on Table 8 (bold shows best result and HMAE results are background colored). As usual, we tested the statistical significance of

Model	BRACS → BACH
HACT-Net	0.402±.028
TransPath	0.618±0.048
TransMIL	0.465±0.100
CLAM-SB/B	0.575±0.036
ScoreNet	**0.734±0.035**
HMAE	0.687±0.032

Table 8: Testing Results on the BACH dataset (weighted F1-score).

the difference in performance between HMAE and the top performer (ScoreNet); Table 9 shows that there is no significant difference between the approches, in terms of weighted F1-score (on 4 classes). We can then conclude that also in this case, HMAE performs at the top level.

Qualitative Evaluation

In order to have a better understanding of the quality of the learned representations in the HMAE model, we have also considered t-SNE dimensionality reduction on the reference

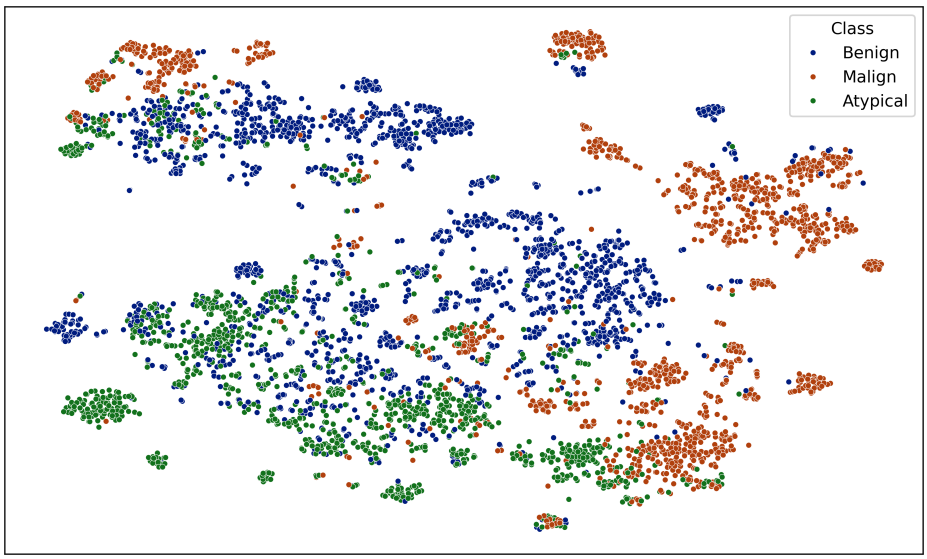

Figure 3: t-SNE visualization of the region of interest (RoI) embeddings from the BRACS dataset. Each point represents a sample, colored according to its class: benign (blue), malignant (red), and atypical (green). This visualization highlights the clustering of different histopathological categories based on their learned feature representations.

	Scorenet	**HMAE**	p-value
WF1-score	0.734	0.687	0.0522

Table 9: BRACS to BACH test: p-value

dataset BRACS and with the higher level of the class hierearchy (benign, cancer and atypical tissue). Figure 3 provides a visualization of the embeddings in a 2D latent space. Each embedding is plotted and colored according to its corresponding class label. Interestingly, even though the model was trained unsupervised, a good degree of class separation is already evident.

Finally, we have analyzed the attention maps, a crucial component of the transformer architecture. Attention maps reveal which parts of the input image the model focuses on the most. By visualizing them, we can understand which image regions are most relevant for the model's predictions. Figure 4 showcases four original images alongside the attention maps generated by the final layer of the MAE in the HMAE architecture. Notably, even during unsupervised training, the model appears to differentiate between connective and glandular tissue. This distinction likely arises because glandular tissue is structurally more complex, requiring the model to retain more information for reconstruction.

Discussion and Future Works

In this work we have presented a masked autoencoder architecture using a Vision Transformer (ViT) as the embedding module, focusing on the representation and classification of histopathological images for breast cancer diagnosis. The model effectively generates informative representations of input images by masking random regions and reconstructing the masked areas in an SSL setting. Applied to histopathological breast cancer images, the model successfully captures relevant features from both tumor and non-tumor regions. These learned representations can then be effectively used as input to a classification model, achieving accurate cancer type and sub-type identification.

We have benchmarked the model with some of the state-of-the-art approaches proposed for the same task and evaluated on the same reference datasets. The HMAE approach exhibits performance at the same accuracy level of the most performing tested techniques. It is worth noting that the latent space of the MAE is only optimized for image reconstruction, and not specifically for classification tasks unlike the benchmark models. This observation strengthens the positive outcomes and underscores the efficacy of employing random input masking. Furthermore, the model's generalization capabilities suggest the potential for being applicable across diverse breast cancer datasets.

Moreover, due to the large spatial resolution of WSIs, data augmentation can be easily adopted. By extracting a larger number of random regions from each WSI, the dataset size can be significantly expanded while minimizing redundancy within the generated images. This step would be important to achieve a more balanced representation of cancerous versus non-cancerous tissues within the training data, without the need for large labeled datasets. Because of the basic peculiarities of masked image modeling, there is a wide expectation to improve the performance of visual models such as HMAE, both in terms of architecture and data scaling, especially in the field of histopathology (Xie et al. 2023; Filiot et al. 2023).

In future works, we plan to explore the impact of expanding the training dataset to enhance the model's ability

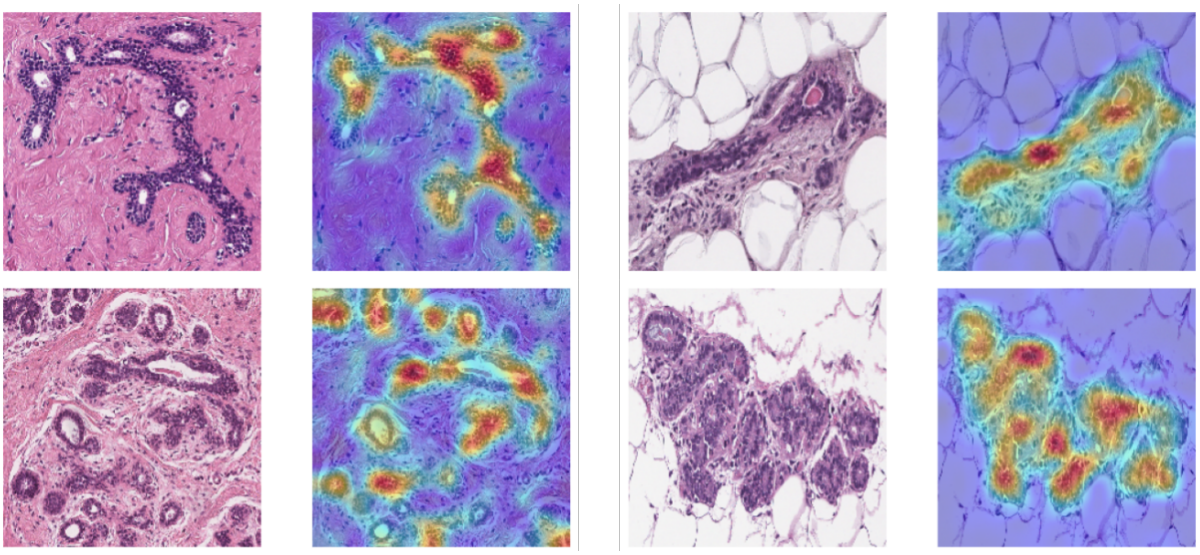

Figure 4: Self-attention heatmaps over histopathology images. The left column in each pair shows the original histopathology image, while the right column presents the corresponding heatmap, where warmer colors (red, yellow) indicate regions of high attention.

to differentiate cancerous and non-cancerous regions. Additionally, to assess the model's generalizability, we propose training it on a collection of diverse datasets. Finally, we also aim to explore more in depth the dataset augmentation given by the random cropping of the WSIs by experimenting with more sophisticated methods that could potentially improve the actual model performance.

Acknowledgments

We acknowledge the use of Chameleon Cloud (Keahey et al. 2020), providing part of the computational resources adopted in the present work. Marco Dossena and Christopher Irwin are PhD students enrolled in the National PhD in Artificial Intelligence for Healthcare and Life Sciences, XXXVIII cycle, Università Campus Bio-Medico, Roma.

References

Aresta, G.; Araújo, T.; Kwok, S.; Chennamsetty, S. S.; Safwan, M.; Alex, V.; Marami, B.; Prastawa, M.; Chan, M.; Donovan, M.; et al. 2019. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56: 122–139.

Aygüneş, B.; Aksoy, S.; Cinbiş, R. G.; Kösemehmetoğlu, K.; Önder, S.; and Üner, A. 2020. Graph convolutional networks for region of interest classification in breast histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, 134–141. SPIE.

Brancati, N.; Anniciello, A. M.; Pati, P.; Riccio, D.; Scognamiglio, G.; Jaume, G.; De Pietro, G.; Di Bonito, M.; Foncubierta, A.; Botti, G.; et al. 2022. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Filiot, A.; Ghermi, R.; Olivier, A.; Jacob, P.; Fidon, L.; Camara, A.; Mac Kain, A.; Saillard, C.; and Schiratti, J.-B. 2023. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023–07.

Gul, A. G.; Cetin, O.; Reich, C.; Flinner, N.; Prangemeier, T.; and Koeppl, H. 2022. Histopathological image classification based on self-supervised vision transformer and weak labels. In *Medical Imaging 2022: Digital and Computational Pathology*, volume 12039, 366–373. SPIE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hou, L.; Samaras, D.; Kurc, T. M.; Gao, Y.; Davis, J. E.; and Saltz, J. H. 2016. Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2424–2433.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Keahey, K.; Anderson, J.; Zhen, Z.; Riteau, P.; Ruth, P.; Stanzione, D.; Cevik, M.; Colleran, J.; Gunawi, H. S.; Hammock, C.; Mambretti, J.; Barnes, A.; Halbach, F.; Rocha, A.; and Stubbs, J. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L. 1989. Handwritten Digit Recognition with a Back-Propagation Network. In Touretzky, D., ed., *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19830–19839.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6): 555–570.

Pati, P.; Jaume, G.; Fernandes, L. A.; Foncubierta-Rodríguez, A.; Feroce, F.; Anniciello, A. M.; Scognamiglio, G.; Brancati, N.; Riccio, D.; Di Bonito, M.; et al. 2020. Hactnet: A hierarchical cell-to-tissue graph neural network for histopathological image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, 208–219. Springer.

Pati, P.; Jaume, G.; Foncubierta-Rodriguez, A.; Feroce, F.; Anniciello, A. M.; Scognamiglio, G.; Brancati, N.; Fiche, M.; Dubruc, E.; Riccio, D.; et al. 2022. Hierarchical graph representations in digital pathology. *Medical image analysis*, 75: 102264.

Qiong, L.; Chaofan, L.; Jinnan, T.; Liping, C.; and Jianxiang, S. 2025. Medical image segmentation based on frequency domain decomposition SVD linear attention. *Scientific Reports*, 15(1): 2833.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Srinidhi, C. L.; Ciga, O.; and Martel, A. L. 2021. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67: 101813.

Stegmüller, T.; Bozorgtabar, B.; Spahr, A.; and Thiran, J.-P. 2023. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. In *Proceedings of the IEEE/CVF winter Conference on applications of computer vision*, 6170–6179.

Tang, W.; Huang, S.; Zhang, X.; Zhou, F.; Zhang, Y.; and Liu, B. 2023. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4078–4087.

Van der Laak, J.; Litjens, G.; and Ciompi, F. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5): 775–784.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J.; Mao, Y.; Guan, N.; and Xue, C. J. 2024. Advances in Multiple Instance Learning for Whole Slide Image Analysis: Techniques, Challenges, and Future Directions. *arXiv preprint arXiv:2408.09476*.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; and Han, X. 2021a. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 186–195. Springer.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; and Han, X. 2021b. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 186–195. Springer.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Wei, Y.; Dai, Q.; and Hu, H. 2023. On data scaling in masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10365—-10374.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18802–18812.

Zhang, Y.; Li, H.; Sun, Y.; Zheng, S.; Zhu, C.; and Yang, L. 2024. Attention-challenging multiple instance learning for whole slide image classification. In *18th European Conference on Computer Vision*, 125–143. Springer.

Zhou, Y.; Graham, S.; Alemi Koohbanani, N.; Shaban, M.; Heng, P.-A.; and Rajpoot, N. 2019. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.