CS512 Class Project Final Report: Exploring masked image modeling-based self-supervised learning method for breast tumor pathological image classification

Zong Fan
zongfan2@illinois.edu
University of Illinois at Urbana, Champaign
Urbana, Illinois, USA

ACM Reference Format:
Zong Fan. 2022. CS512 Class Project Final Report: Exploring masked image modeling-based self-supervised learning method for breast tumor pathological image classification. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Digital computational pathological imaging is a useful method to quantify lesion status for early breast tumor diagnoses like tumor staging and risk assessment. Recently, deep learning-based methods have achieved great success in the task of differentiating tumor types accurately. However, most of these methods are fully-supervised methods, which means that accurate classification relies on large amounts of accurate fully-annotated training data. To reduce the burden of large-scale annotation, we introduced a masked image modeling-based self-supervised learning algorithm, SimMIM, to learn meaningful feature representations from unlabeled images. The extracted representations render high performance to be used for downstream classification. Experimental results based on a public breast cancer pathology image dataset show similar or even higher classification performance compared to a conventional fully-supervised classification method. In addition, we find that SimMIM learns better representations by use of random masking of the input image with a larger masked patch size as 64×64 pixels, which is different from the original paper. Using Swin Transformer as the encoder, our approach achieves 84.53% top-1 fine-tuning classification accuracy, surpassing the conventional ResNet50 by 0.9%.

2 INTRODUCTION

Breast cancer, the most common neoplasm diagnosed in women, is the most frequent cause of death in women between 35-55 years of age [26, 23]. American Cancer Society suggested that women aged 45 to 54 years should get mammograms every year [24]. However, the sensitivity of mammography is relatively low in women with dense breasts [23, 17]. This is an important limitation because about 40%-50% women in America aged 40-74 have dense breast [14], and breast density itself has been shown as a breast cancer risk factor [12, 3, 22].

With the development of electrical microscopes, acquired whole-slide images (WSI) provide rich texture and pattern information of lesion cells of breast tumors. This can support the analysis of tumor samples and help pathologists make clinical decisions on lesion diagnosis, treatment strategy design, and other clinical applications. Recently, several deep learning (DL)-based methods have been applied to achieve this goal of automatic discrimination by using pathology images. CSCDNN, similar to a typical deep convolutional neural network AlexNet [16] that can handle up to 1000 categories of classification, has shown good performance in classifying breast tumor types with pathology images [11]. An optimized hierarchical network called BiCNN was proposed by Wei et.al to handle the problem of subtle differences of inter-class histopathological images [37].

While the automated multiple lesion type classification has promising future applications, some practical challenges need to be solved. particularly, supervised learning might suffer from generalization error, spurious correlations, and adversarial attacks [19]. The first challenge is the limited number of breast tumor samples. Unlike large-scale natural image dataset like ImageNet [27], the breast tumor pathological image datasets only have hundreds of WSI images, such as TCGA [38]. Annotating data can be very time- and effort-consuming, especially when annotating process need expertise knowledge which can be very costly. Even though image patches can be extracted from the initial high-resolution images, it's still quite difficult to make the model learn the data distribution decently without the disturbance of overfitting. It's highly likely that the model memorizes the characteristics of the specific small dataset rather than the intrinsic image-related features. The second challenge is related to the training dataset imbalance. Generally, from the level of subordinate types of specific cancer, the distribution of each type can be very imbalanced. For example, ductal carcinoma is the most common breast cancer subtype, as it accounts for 80% of all breast cancer, while only about 10% of diagnosed invasive breast cancers are lobular carcinoma [29]. In addition, the overall ratio between the number of benign images and malignant images is about one-third. Another challenge comes from the intrinsic properties of pathology images that lesions from different types show a high similarity of texture patterns on histopathology images, which places high demands on methods to effectively detect subtle texture differences. These problems make us unable to ignore the risks overfitting, which may degrade the model performance on unseen examples.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To alleviate the problem of obtaining large-scale annotated image datasets, a lot of methods have been proposed. Transfer learning, where training on a widely different dataset, can reduce the amount of training data required in a target task [42]. Semi-supervised learning is another popular solution that focuses on automatically exploiting incomplete or inexact supervisions (or unlabeled data) to improve model performance. For example, Lie et.al employed a semi-supervised learning algorithm to leverage the unlabled images by pseudo labeling to improve the classification performance and reduce the number of annotated images [18]. As a promising alternative, self-supervised learning has recently emerged as an important and powerful technique to create a pre-trained model for downstream applications without the need for large-scale annotated datasets [30, 19, 39]. Pre-training is conducted on unlabeled data based on a defined proxy objective, which enables computers to label, categorize, and analyze data automatically [39]. The intuition of self-supervised learning is to leverage the inherent co-occurrence relationships as the self-supervision from large amounts of unlabeled data [19]. Such inherent knowledge can help the model learn the data distribution more comprehensively, significantly alleviating the out-of-distribution or generalization problem and improving the model performance [41]. Ideally, the model can learn general understanding of the image content and input data distribution.

In this paper, we investigate how a self-supervised learning method called SimMIM based on masked modeling principle [40] can be extended to learn informative representations for breast tumor pathological image classification. Briefly, SimMIM masks random patches from the input image and reconstructs the missing patches in the pixel space by use of transformer [34] architecture via regression, where the transformer plays the role of an encoder to learn representations from local image patches. A previous self-supervised method developed for histopathology showed that using transformer architectures to capture global information can help extract better feature representations [36]. Therefore, in this study, we consider pathological images, and compare the outcome between a conventional fully-supervised learning method and our method, as well as learn how the different mask settings of the framework contribute to the learning outcome. A medical image dataset with breast tumor histopathological images called BreakHis [29] was also employed for experimenting and studying these ideas. Using Swin Transformer [20] as the encoder, our approach achieves 84.53% top-1 fine-tuning classification accuracy, surpassing the conventional ResNet50 by 0.9%.

The rest of the paper is organized as follows: In Section 2, we discuss the related work about self-supervised learning and transformer. In Section 3, we briefly explain the details of our methods and the network architectures. The model training and implementation details are shown in Section 4. The experimental results are presented in Section 5. Finally, in Section 6, we make a conclusion and discussion with an outlook on what needs to be considered for further improving self-supervised learning in the medical pathological imaging field.

3 RELATED WORKS

3.1 Self-supervised learning

Considering the difference in model architecture and objectives, most self-supervised learning methods can be grouped into three types: contrastive methods, generative methods, and adversarial methods [19].

Contrastive methods would train an encoder to encode the input sample into an explicit vector to measure feature similarity. Significant attention has been given to instance discrimination through contrastive learning with multiple views [9]. For example, Bachman et al. proposed a contrastive self-supervised method based on creating multiple views using augmentation [1]. Other recently proposed constrastive self-supervised methods like CMC [31] and SimCLR [5] have achieved the performance with the reduced gap between supervised and unsupervised training on ImageNet Dataset.

In order to simulate input data distribution, generative methods train an encoder to encode the input sample into an explicit vector and a decoder to reconstruct the sample from the vector. Popular generative-based self-supervised learning methods include auto-regressive models, flow-based models, and auto-encoding models [19]. For example, Oord et al. proposed PixelCNN to model images pixel by pixel via auto-regressive method [32]. Flow-based models generate the target complicated densities step by step via stacking a series of transforming functions that describe different data characteristics, such as NICE algorithm proposed by Dinh et al. [7]. Variational auto-encoding and its variants follow the assumption that data are generated from underlying latent representation and this posterior distribution can be approximated by a variational distribution given some data samples [33].

Adversarial methods employ encoder-decoder architectures to generate fake samples and a discriminator to distinguish fake samples from real samples. The previous studies show that the performance of generative methods is less competitive than the other two methods, primarily due to the inherent defects of the point-wise nature of the generative objective [19]. Generative adversarial network (GAN) and its variants are widely-used for representation learning in this adversarial fashion. For example, BiGAN chooses to extract the implicit distribution directly by reconstructing the whole input [8]. Inpainting is another way which asks the model to predict an arbitrary part of an image given the rest of it [25]. A previous study found that image inpainting might be a strong self-supervised pre-text task, stronger inpainting capability does not necessarily lead to stronger fine-tuning performance on downstream tasks [40].

Particularly, the masked image modeling method used in this study is a hybrid method, which masks a rectangle area of the original images and predicts the missing pixels via a verification task with a contrastive predictive coding loss [13].

3.2 Transformer architecture

Transformers architecture, which has achieved great success in the natural language processing field, is also introduced in the computer vision field. The transformer's self-attention layers can complement classical CNN network backbone or heads by providing the capability to encode distant dependencies [35, 15]. Also, the

encoder-decoder design of the transformer has been applied for the object detection [4] and segmentation tasks [6].

Vision Transformer (ViT) is the pioneering work to directly applies a Transformer architecture on exclusive and ordered image patches for image classification [10]. It has an impressive speed-accuracy tradeoff on image classification compared to conventional CNN networks. Some works apply ViT models to the dense vision tasks of object detection and semantic segmentation by direct upsampling or deconvolution but with relatively lower performance [2, 43]. In this study, Swin Transformer [20] architecture is employed, which is revised based on ViT architecture. It achieves the best speed-accuracy trade-off among these methods on image classification. Specifically, Swin Transformer modified the self-attention computation block for each patch token and proposed shifted window partitioning strategy to introduce connections between neighboring non-overlapping windows in the previous layer to improve modeling power.

4 METHODS

Although the idea of masked autoencoding in NLP field has attracted a lot of research interest, the research progress of this technique in computer vision still lags behind NLP, which may be explained by the difficulties in transferring the method to different data modalities. In this study, we mainly explore the SimMIM method as the self-supervised representation learning method and investigate its effect on computer vision tasks, especially the classification performance on unlabeled data samples [40]. SimMIM was proposed by Xie et al [40], which masks random patches from the input image and reconstructs the missing patches in the pixel space by use of transformer architecture via regression. As shown in Figure. 1, SimMIM framework has four major components as discussed below.

4.1 Masking strategy

The first step in the framework is employing a masking strategy to mask a portion of an input image. This module designs how to select the area to mask, and how to implement masking of the selected area. The masked image is used as the input for the following transformer encoder. A learnable mask token vector to replace each masked patch was used. The dimension of token vector is the same as that of the other visible patch representation after patch embedding. Patch-aligned random masking strategy was used in this study to select the masking area. As shown in Figure. 2, each image patch is a basic processing unit for the following Transformer encoder, either fully visible or fully masked. In this study, we employed equivalent patch sizes of two different resolution stages, 32×32 to 64×64.

4.2 Transformer encoder

In this study, the encoder of Swin Transformer [20] is employed as the encoder for the masked image patches to extract the image representations. The overall Swin Transformer architecture is shown in Figure 3. As shown in Figure **??**, after the input image is split into non-overlapping image patches by given masking strategy, each image patch is treated as a unit token and its representation is obtained by concatenating the raw pixel values in RGB format. A patch size of 4×4 was employed in this study, so the feature dimension is 48 ($4 \times 4 \times 3$). To achieve hierarchical representation, the number of tokens is reduced as the network gets deeper via patch merging. Particularly, Swin Transformer block shown in Figure 3 (b) modified the standard multi-head self-attention module with shifted windows to reduce the global computation complexity and gain connects across different windows. Each Swin Transformer block has a shifted window-based multi-head self-attention module, followed by a 2-layer multi-linear perceptron inter-connected by GELU nonlinearity activation layer. A LayerNorm layer is applied before each multi-head self-attention module and each multi-linear perceptron. A residual connection is applied after each block.

4.3 Prediction head

After the representation learning in the Swin Transformer encoder, a prediction head is applied to predict the target. It can project the learned representation into an arbitrary dimension. According to the Swin-Transformer paper, the prediction head can be made extremely light weight with enough capacity and performance [20]. Therefore, in this study, two fully-connected layers with 1024 neurons were used.

4.4 Prediction target

Considering the input images are unlabeled, a straight-forward option for the prediction target is to predict raw pixels of the masked area. To predict the pixel values in the exact size of input images, each representation is mapped back to the original input image size to predict the corresponding raw pixels. L1-loss is employed to evaluate the quality of reconstructed mask pixel values.

5 IMPLEMENTAION DETAILS

5.0.1 Datasets. The public breast tumor histopathological image database, BreaKHis (http://web.inf.ufpr.br/vri/breast-cancer-database), is utilized to demonstrate the performance of the selected method. The whole-slide images of HE-strained histopathological images provide rich texture and pattern information of breast tumor lesions. This can support the analysis of tumor samples and help pathologists make clinical decisions on lesion diagnosis, treatment strategy design, and other clinical applications. From the perspective of clinical diagnosis, digital pathology images can be employed to differentiate lesion subtypes or binary classification between malignancy and benign. However, annotating the histopathological images usually needs expertise knowledge and large amounts of time, which is very expensive and time-consuming in obtaining a large-scale fully annotated dataset. Therefore, it's always desired to utilize the unlabeled image to facilitate prediction accuracy.

This database includes a total of 7909 breast cancer histopathology images acquired on 82 different patients. In this database, four benign breast tumor subtypes: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant breast tumor subtypes: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC) are available. Also, these images have four kinds of magnifications, including 40X, 100X, 200X, and 400X. In this study, 400X images would be first investigated since 400X images usually achieved better performance using the supervised learning method [29]. The

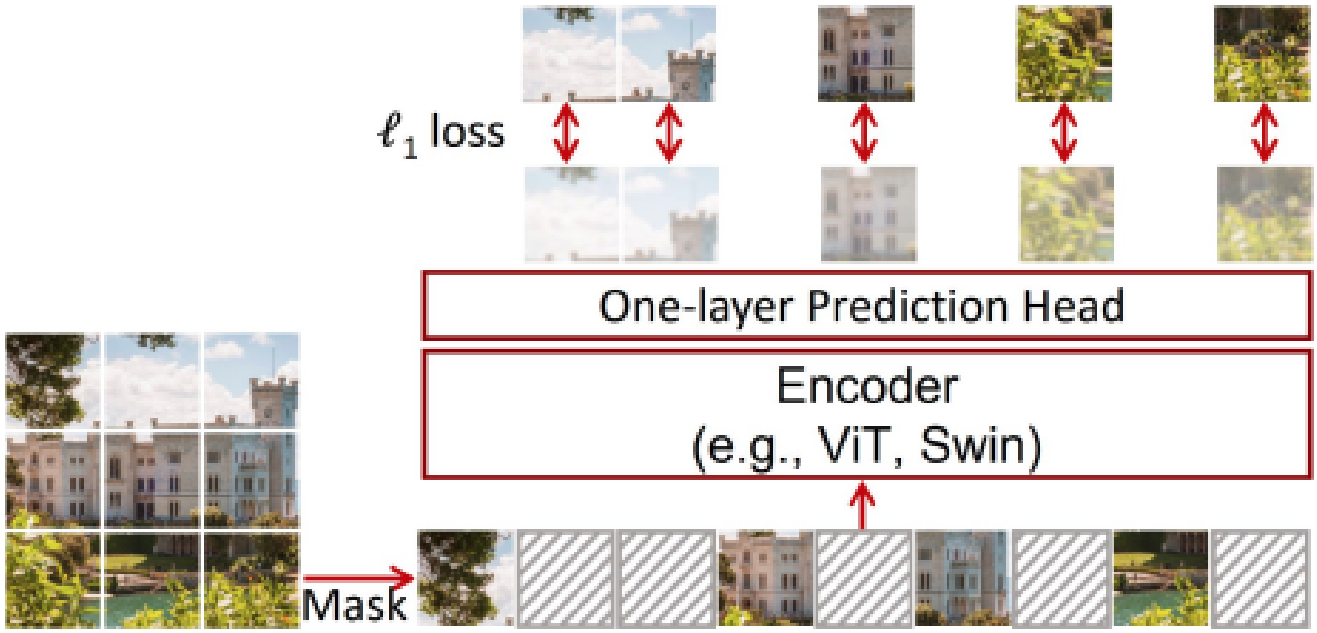

Figure 1: Framework architecture of SimMIM. After the encoder extracts representations of each patch, a one-layer lightweight head is designed to predict the raw pixel values of the randomly masked patches. A simple L1 loss was employed to supervise the model training. Credit: Xie et al. 2022 [40]

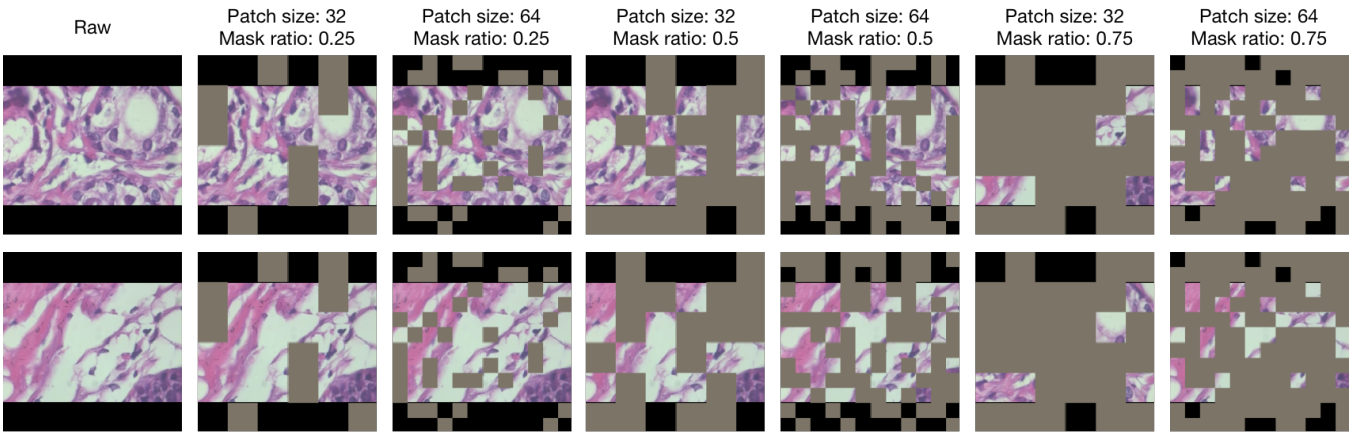

Figure 2: Illustration of masking area generated by random masking strategy using different mask ratios ranging from 0.25 to 0.75.

detail of the 400x BreakHis data used in experiments is shown in Table 1 and several image samples are shown in Figure 4.

5.0.2 Model training. To prepare the training and testing data, a set of images containing 10% examples from each class label was selected as the testing dataset. The rest images were randomly divided into the training dataset (85%) and validation dataset (15%).

Before training, model parameters pre-trained on ImageNet dataset [27] were used to initialize the Swin Transformer encoder. This transfer learning strategy is employed to accelerate the training process and improve learning efficiency. In each training iteration, a mini-batch of 4 images were sampled randomly from the training dataset. The images were first padded by 120 pixels on the top and bottom margins to make the input image square without distortion. Then the padded images were preprocessed with the following data augmentation strategies before inputting into the network. The data augmentation included random horizontal flip,

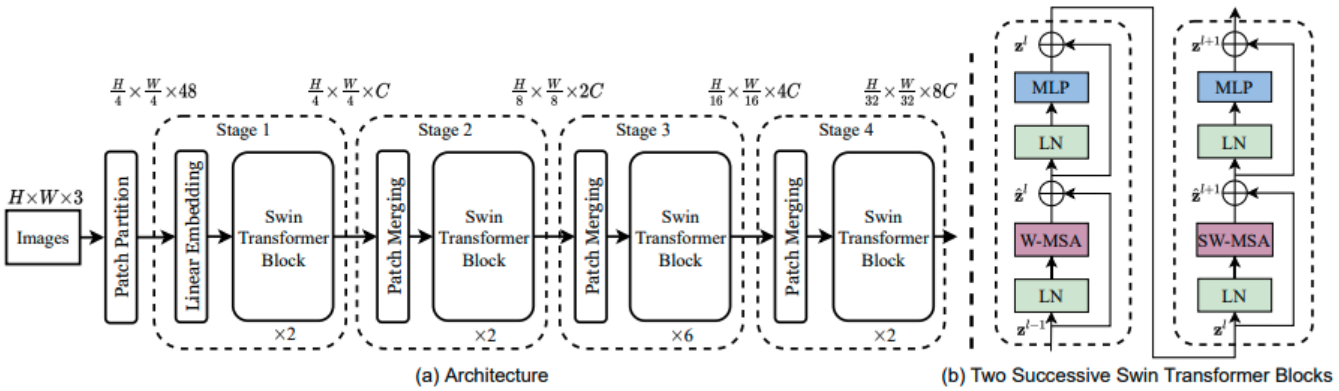

Figure 3: (a) The architecture of a Swin Transformer; (b) two successive Swin Transformer Blocks. Image credits: Liu et al.(2021) [20].

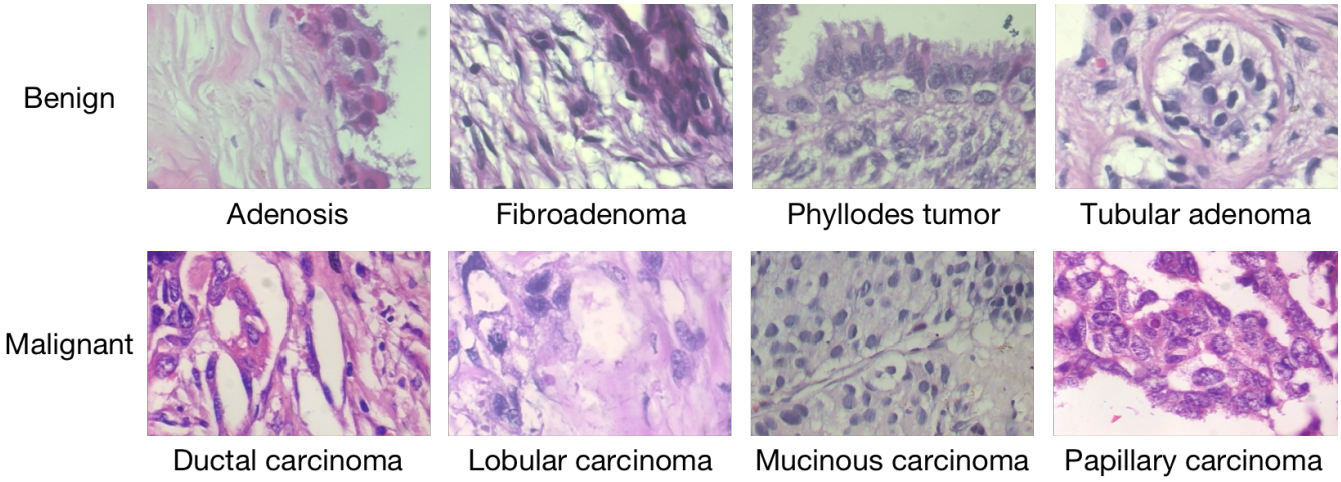

Figure 4: Examples of 400x BresKHis histopathological images

Tumor	type	Number of images
Adenosis	benign	114
Fibroadenoma	benign	253
Phyllodes tumor	benign	109
Tubular adenoma	benign	149
Ductal carcinoma	malignant	864
Lobular carcinoma	malignant	156
Mucinous carcinoma	malignancy	205
Papillary carcinoma	malignancy	145

Table 1: Distribution of 400X images

random vertical flip, and random center crop. The preprocessed images were resized to 384×384 pixels by bi-linear interpolation. Finally, normalize the pixel values following the ImageNet normalization scheme [27]. The processed images were the input images for the masking module. In this study, patch sizes of 32×32 and 64×64 were used to investigate how the masking patch size would affect the learned representation quality. The mask ratio was set as 0.5 for all experiments. In the next step, the numbers of Swin Transformer blocks in the four stages were set as 2, 2, 18, and 2, respectively. The numbers of heads in each multi-head attention module were 4, 8, 16, and 32. With L1-loss calculated to measure the reconstruction quality of the masked area, Adam stochastic gradient algorithm [**Kingma2015AdamAM**] was employed as the optimization algorithm, with decay rate $\beta_1 = 0.9, beta_2 = 0.999$ and initial learning rate $lr = 0.0005$. Cosine learning rate strategy was employed to automatically adjust the learning rate according to the training progress [21]. The framework was trained on the training dataset for 500 epochs.

The proposed framework was implemented by use of PyTorch 1.7.0 and was performed training and validation on Nvidia GeForce GTX 1080ti GPUs. The code of SimMIM framework has been released publicly on GitHub (https://github.com/microsoft/SimMIM). We wrote our data preprocessing and model training code to fit the target dataset. The details of training and evaluation are shown in

the appendix scripts. Noticeably, our code relies on the Swin Transformer repository (https://github.com/microsoft/Swin-Transformer).

5.1 Methods for comparison

After the model is trained via the self-supervised learning method, the representation extracted from a given image was used for a downstream classification task. For comparison, the model trained with a ResNet50 in fully-supervised learning was used for comparison.

6 RESULTS

6.1 Classification performance

After training the Swin Transformer encoder on the whole dataset without using the class labels, the images in the training dataset were first used to extract their corresponding feature representations. Then these representations were used to train the prediction head using the class labels in a fully-supervised manner. For comparison, we trained a ResNet50 classification model using the same training dataset. First, we ran binary classification to classify breast tumor types into benign or malignant. As shown in Figure 5, using SimMIM can always achieve better performance than ResNet50 in terms of accuracy, F1-score, and area under the curve (AUC). Similarly, when considering the multi-class classification task that aims to distinguish which sub-category of the input image belongs, the result shows the difference between SimMIM and ResNet50 is trivial, as shown in Figure 6. These results show that the representation learned via SimiMIM can achieve promising performance in the downstream classification tasks.

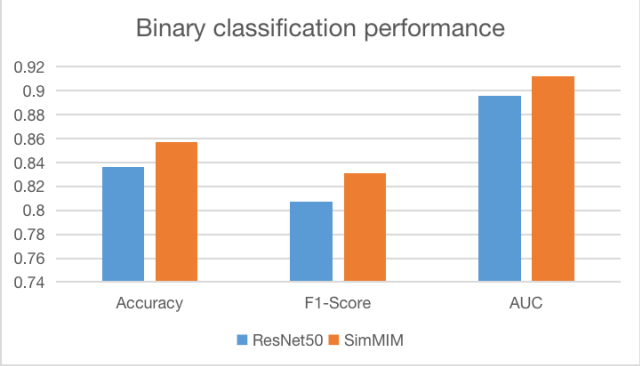

Figure 5: Binary classification (Benign vs. Malignant) performance using SimMIM and ResNet50.

6.2 Representation comparison

One benefit of self-supervised representation learning is that the learned representation is informative and robust, even without particular supervision signals provided. This is why self-supervised learning can reduce the amount of annotated images. To investigate the informativeness of the representation extracted by the SimMIM, we employed K-nearest neighboring classification by using the learned representation without re-training. Also, different K-values and patch sizes were explored to see how the settings would affect

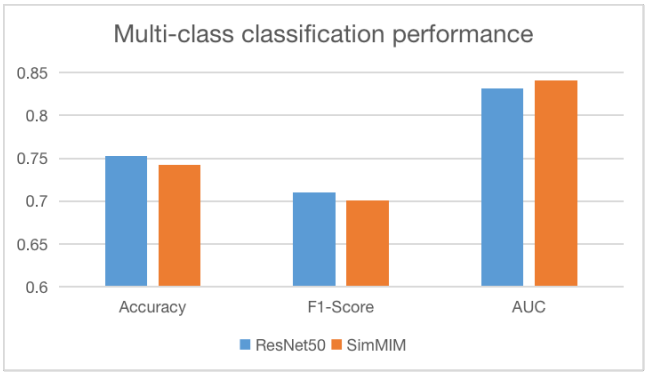

Figure 6: Multi-class classification (8 sub-categories) performance using SimMIM and ResNet50.

the model performance. As shown in Figure 7, using SimMIM with patch size of 64×64 and $K = 74$ achieves the best prediction accuracy on the testing dataset, even higher than the representation learned in the fully supervised learning manner. It indicates that the representation learned by SimMIM has excellent generality to be employed in various downstream applications. Interestingly, using a larger masked patch size seems to work better in this case, implying that breast tumor classification might rely more on local texture and interactions inside each patch rather than the close connections between patches.

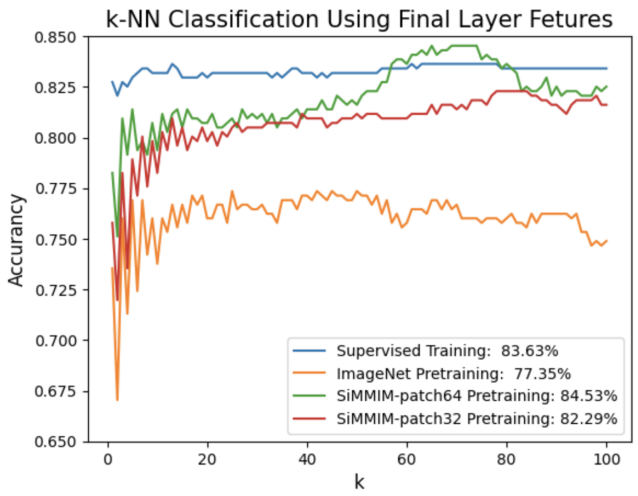

Figure 7: K-nearest neighboring classification using representation from SimMIM and ResNet50. The x-axis shows different K values and y-axis shows the corresponding accuracy at a particular K. Supervised training means using ResNet50 and retraining it on the breast tumor dataset. ImageNet pertaining means using ResNet50 without retraining.

6.3 Feature interpretability

To improve model interpretability, class activation mapping (CAM) was proposed to describe how a DL-based model predicts the outcomes by identifying discriminative regions on the given image and is employed broadly to explain classification networks. Particularly, Grad-CAM, an effective generalization of CAM which is applicable for a variety of CNN models, can provide class-discriminative localization for visual explanations [28]. It computes the gradient score for each class in terms of the feature map activations of the last convolutional layer in the CNN model which renders high-level semantic and spatial information. In this study, Grad-CAM was employed to highlight the potential ROIs based on the extracted feature of the last convolutional layer of ResNet50, representing the spatial attention learned by the model for classification. In addition, the attention map learned in the multi-head attention module of the last Swin Transformer block is visualized to check the area focused during the self-learning process. As shown in Figure 8, the saliency achieved by the attention learned by the transformer encoder seems to be more reasonable than that achieved by the Grad-CAM. The attention regions are more focused and match better to the lesion location. The result shows the informativeness of learned representation using the SimMIM self-learning framework, which has high efficiency in extracting essential information even without subversion.

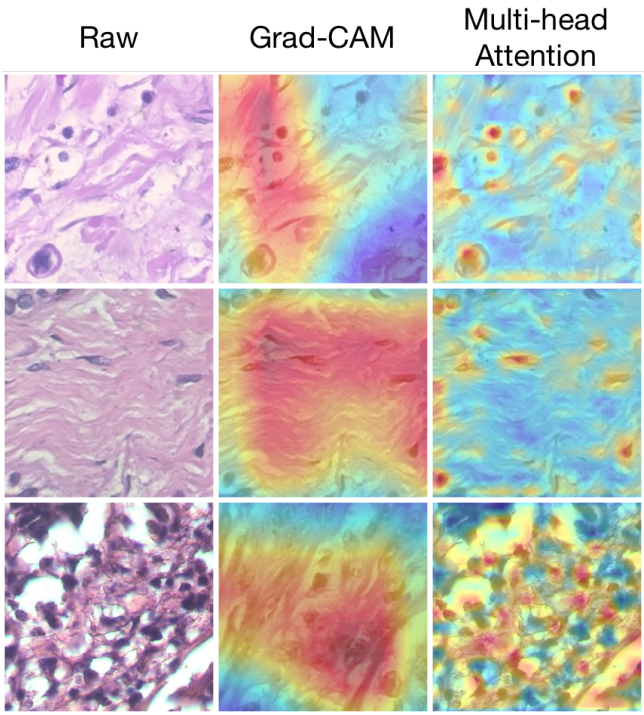

Figure 8: Saliency map visualization for the last convolutional layer of ResNet50 vs the attention learned in the last Swin Transformer encoder. The reddish area indicates where the learned representation focuses.

7 DISCUSSION

Breast cancer pathology images are utilized as the example dataset to show the SimMIM's performance. The results show that SimMIM might achieve superior classification performance compared to conventional fully-supervised learning methods. In the experiments, pathology images of breast tumor samples with 400x magnification pathology images are used as input images, showing a promising result on pathology images both binary and multi-classification.

To understand the philosophy of framework design, it's important to know the difference between the task of natural language processing and computer vision. First, the information density is different between language and vision. Images are raw and low-level signals which have high spatial redundancy and relatively low information density. Language is high-level human-generated data that is highly semantic and has high information density. The visual signals are more continuous than the discretization of language tokens. Second, the images have a strong locality where neighboring pixels usually share highly-correlated information, while the language tokens don't show such a strong locality. Third, the decoder of the autoencoder architecture maps the latent representation back to the input, which is important for reconstructing texts and images. In the vision field, the decoder reconstructs pixels that are of a lower semantic level. In contrast, the decoder predicts missing words that contain rich semantic information in language. Therefore, SiMMIM proposed several innovative designs to address these differences as follows. First, random masking strategy which makes the input sparse for the encoder to efficiently learn the meaningful representation from the input image for reconstruction; a raw pixel regression task was designed to predict the value of missing pixels. This design aligns well with the continuous nature of visual signals. It was conducted by a lightweight prediction head which can significantly speed up training.

One future research directive is to investigate the impacts of the levels of image magnification on classification performance. We will also investigate its performance on other tumor types, such as cervical cancer and head neck cancer. Also, other popular self-learning methods such as contrastive learning-based method would be investigated.

Since future therapeutic concepts in breast cancer are aimed at the individualization of therapy and treatment escalation and de-escalation based on tumor biology and early therapy response9, our method can perform automated breast cancer multi-classification from digital pathological images and has excellent potential to be applied in clinical settings.

REFERENCES

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

[2] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. 2020. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*.

[3] Norman F Boyd et al. 2007. Mammographic density and the risk and detection of breast cancer. *New England journal of medicine*, 356, 3, 227–236.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] Cheng Chi, Fangyun Wei, and Han Hu. 2020. Relationnet++: bridging visual representations for object detection via transformer decoder. *Advances in Neural Information Processing Systems*, 33, 13564–13574.

[7] Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

[9] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

[10] Alexey Dosovitskiy et al. 2020. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[11] Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li. 2017. Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7, 1, 1–10.

[12] Jennifer A Harvey and Viktor E Bovbjerg. 2004. Quantitative assessment of mammographic breast density: relationship with breast cancer risk. *Radiology*, 230, 1, 29–41.

[13] Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*. PMLR, 4182–4192.

[14] N Howlader, AM Noone, M Krapcho, D Miller, K Bishop, CL Kosary, et al. 2018. American cancer society cancer facts & figures 2018. *Atlanta: American Cancer Society*.

[15] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. 2019. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3464–3473.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 6, 84–90.

[17] Carol H Lee et al. 2010. Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *Journal of the American college of radiology*, 7, 1, 18–27.

[18] Kun Liu, Zhuolin Liu, and Sidong Liu. 2022. Semi-supervised breast histopathological image classification with self-training based on non-linear distance metric. *IET Image Processing*, 16, 12, 3164–3176.

[19] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. 2020. Self-supervised learning: generative or contrastive. *ArXiv*, abs/2006.08218.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

[21] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

[22] Valerie A McCormack and Isabel dos Santos Silva. 2006. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology and Prevention Biomarkers*, 15, 6, 1159–1169.

[23] Monika Nothacker, Volker Duda, Markus Hahn, Mathias Warm, Friedrich Degenhardt, Helmut Madjar, Susanne Weinbrenner, and Ute-Susann Albert. 2009. Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. a systematic review. *BMC cancer*, 9, 1, 1–9.

[24] Kevin C Oeffinger et al. 2015. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama*, 314, 15, 1599–1614.

[25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

[26] Devish Pyakurel, S Karki, and CS Agrawal. 2014. A study on microvascular density in breast carcinoma. *Journal of Pathology of Nepal*, 4, 7, 570–575.

[27] Olga Russakovsky et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 3, 211–252.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

[29] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. 2015. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63, 7, 1455–1462.

[30] Karin Stacke, Jonas Unger, Claes Lundström, and Gabriel Eilertsen. 2021. Learning representations with contrastive self-supervised learning for histopathology applications. *arXiv preprint arXiv:2112.05760*.

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European conference on computer vision*. Springer, 776–794.

[32] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

[33] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

[36] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. 2021. Transpath: transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 186–195.

[37] Benzheng Wei, Zhongyi Han, Xueying He, and Yilong Yin. 2017. Deep learning model based breast cancer histopathological image classification. In *2017 IEEE 2nd international conference on cloud computing and big data analysis (ICCCBDA)*. IEEE, 348–353.

[38] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45, 10, 1113–1120.

[39] Lirong Wu, Haitao Lin, Zhangyang Gao, Cheng Tan, and Stan.Z.Li. 2021. Self-supervised learning on graphs: contrastive, generative,or predictive. *IEEE Transactions on Knowledge and Data Engineering*.

[40] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: a simple framework for masked image modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9643–9653.

[41] Keyulu Xu, Jingling Li, Mozhi Zhang, Simon Shaolei Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2021. How neural networks extrapolate: from feedforward to graph neural networks. *ArXiv*, abs/2009.11848.

[42] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

[43] Sixiao Zheng et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Roles: Zong Fan is responding to all literature searching, coding, model training, and paper writing.