Traitement du Signal
Vol. 41, No. 2, April, 2024, pp. 729-738
Journal homepage: http://iieta.org/journals/ts

Enhancing Lung Cancer Diagnosis with MixMAE: Integrating Mixup and Masked Autoencoders for Superior Pathological Image Analysis

Bin Zuo*, Feng Xiong

Beijing Institute of Remote Sensing Information, Beijing 100011, China

Corresponding Author Email: zuobin97117@163.com

https://doi.org/10.18280/ts.410215

Received: 17 November 2023
Revised: 2 February 2024
Accepted: 26 February 2024
Available online: 30 April 2024

Keywords:
self-supervised learning (SSL), mask autoencoders (MAE), Mixup, image classification, lung cancer

ABSTRACT

Lung cancer, standing as the world's second most fatal ailment, inflicts profound and irreversible damage on human life. Histopathology, the microscopic examination of tissues, is pivotal for the accurate diagnosis and effective treatment of this malignancy. Yet, the burgeoning volume of lung cancer pathological images and their inherent complexity present formidable challenges within the diagnostic landscape. In response, we introduce a novel hybrid methodology, Mixup Masked Autoencoders (MixMAE), marrying the Masked Autoencoder (MAE) with the image Mixup technique, rooted in prior pathological insights, to discern lung cancer pathological images with heightened acuity. Leveraging self-supervised learning (SSL) models, MixMAE enhances the precision of lung cancer treatment by infusing Mixup designs into MAE's upstream tasks. This process involves feeding Mixup lung cancer images into MAE, enabling the model to capture an enriched tapestry of lung cancer image features within the constrained visibility afforded by a high mask rate, thus elevating learning efficiency. To corroborate the logic and efficacy of our model, we curated a dataset of 7,062 lung cancer pathological images for experimentation. Incorporating the Mixup algorithm into MAE significantly uplifted the diagnostic accuracy to 95.64%, surpassing the original MAE model in classification efficacy. Moreover, acknowledging clinical imperatives, we assessed our model's generalization capacity against the LC25000 public dataset, a compendium of vastly differing data volumes and categories. These experiments affirm MixMAE's adeptness not only in identifying lung cancer pathological images but also in distinguishing other cancer types with superior accuracy relative to other complexly engineered networks.

1. INTRODUCTION

Lung cancer remains the leading cause of cancer-related morbidity and mortality worldwide [1]. Key challenges in addressing this issue include the inadequate detection of cancer, the scarcity of effective treatments, and suboptimal treatment outcomes. However, early intervention significantly enhances survival rates. Histopathological examination is the clinical gold standard for lung cancer screening, providing detailed insights into tissue morphology and tumor subtypes through the analysis of biopsy and resection samples [2]. Despite the clarity and detail offered by histopathological images, their interpretation by pathologists is time-consuming and subject to variability.

The advent of computer-aided pathological image detection technology, propelled by early machine learning and subsequently by advanced deep learning techniques, has marked a significant leap forward [3]. Traditional models like Support Vector Machines (SVM) and Random Forests (RF) [4-9] have given way to deep learning networks, particularly Convolutional Neural Networks (CNNs) [10] and, more recently, Vision Transformers (ViTs), which have demonstrated superior performance in pathology image analysis [11].

Deep learning methods, however, rely heavily on extensive labeled datasets, which are costly and labor-intensive to compile, especially for histopathology. SSL is a machine learning method that bridges the gap between supervised and unsupervised learning. Instead of relying on labeled data in the traditional sense, SSL generates supervised signals from the input data itself to train the model. This approach allows the model to learn a useful representation of the data without explicit labeling, thus leveraging the large amount of unlabeled data to improve learning efficiency and performance. SSL [12, 13] emerges as a promising alternative, leveraging unlabeled data to learn meaningful representations and showing success in various medical imaging tasks, including classification [14], detection [15], and segmentation [16].

Despite progress, SSL methods face challenges in histopathology, struggling to distinguish subtle nuances and extract higher-level semantic information. Addressing this gap, we introduce MixMAE [17], a novel SSL approach for classifying Hematoxylin and Eosin-stained histopathology images. MixMAE innovatively combines Mixup image augmentation with MAE learning, facilitating the extraction of complex features from limited labeled data. The specific structure is shown in Figure 1. The proposed MixMAE extends MAE, and performs pixel-level image Mixup before the mask step in the pre-training stage, so that the model can learn more image information in limited visible blocks.

Our contributions are threefold: MixMAE integrates the strengths of MAE and Mixup, enhancing feature learning for downstream tasks; it excels in analyzing challenging datasets, outperforming existing methods; and in a lung cancer dataset, it effectively differentiates between infiltration types, showcasing its clinical applicability. Despite the thoroughness of our study, the manuscript requires language simplification for improved readability, without altering the core premises.

The structure of this paper is as follows: Section 1 introduces the research background and significance, Section 2 reviews the status of related work in the past few years, Section 3 introduces the method of this paper in detail, Section 4 introduces the experimental data, and Section 5 introduces the experimental setup and evaluation indicators are described. Section 6 introduces the reasons for the analysis of the experimental results and concludes the paper with a brief conclusion.

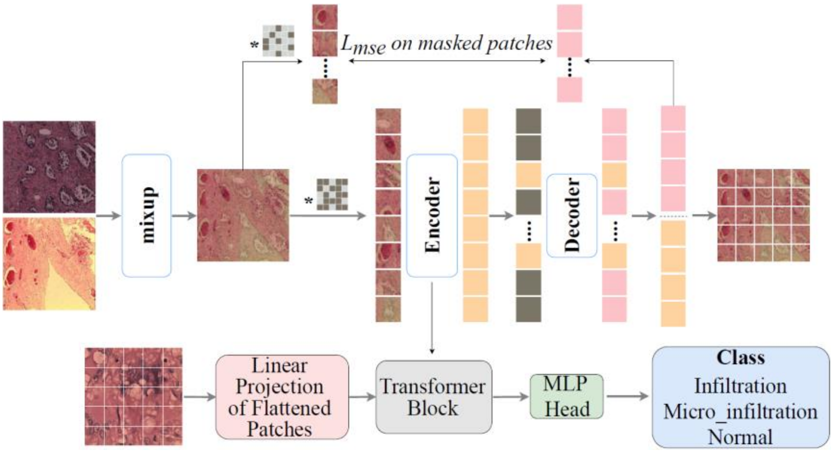


Figure 1. Illustration of the proposed MixMAE method

2. RELATED WORK

The advent of SSL represents a significant evolution within the deep learning domain, enabling models to be effectively trained on unlabeled data. This advancement facilitates the straightforward extraction of visual representations from such models. This section provides an overview of existing techniques for classifying lung cancer through histopathological images, followed by an examination of research into pathology data augmentation and SSL approaches in medical imaging.

2.1 Lung cancer pathological images classification

Deep learning technologies have found significant applications in the pathological analysis of lung cancer, particularly in the realms of early tumor screening and the differentiation of benign and malignant neoplasms. Zhang et al. [18] pioneered the "Early Computer Diagnosis System for Lung Cancer," enabling the detection of diverse lung cancer types through pathological section analysis. Yang et al. [19] introduced a novel six-class classifier leveraging the EfficientNet-B5 model, achieving an intricate multi-class tissue classification that mirrors the complexities of real-world histopathological environments. Utilizing transfer learning and weakly supervised approaches, Kanavati et al. [20] employed a Convolutional Neural Network (CNN) based on the EfficientNet-B3 architecture, trained on a dataset of 3554 whole slide images (WSIs), to discern lung cancer from non-cancerous tissues with notable precision. Furthermore, Chen et al. [21] devised a detection model for lung cancer cells employing both CNN and Swin Transformer, demonstrating not only a reduction in computational demand but also surpassing the performance of the classical CNN model, ResNet50.

These studies underscore the nascent yet evolving state of lung cancer cell detection technologies, which currently suffer from suboptimal accuracy. Traditional CNNs are constrained to extracting localized features via convolutional kernels; in contrast, SSL models, through their utilization of attention mechanisms, are capable of assimilating features from entire images, offering a more nuanced analysis. This comparison indicates a potential paradigm shift towards SSL models for enhanced image analysis in lung cancer detection, embodying a leap towards precision diagnostics.

2.2 Pathological data augmentation

As we all know, pathological images contain higher-level information than ordinary images, and the enhancement quality of pathological images directly determines the next step of clinical diagnosis. In order to provide doctors with clearer and more accurate medical images, and provide good support for subsequent high-level processing such as medical image processing and analysis, image enhancement has been extensively studied in the field of computer pathology. The most commonly used data augmentation methods include mirroring, flipping up and down, scaling and rotation [22, 23], which can improve the robustness of the model and improve the recognition efficiency of the model.

In addition, for the problem of data imbalance, generative models such as Generate Adversarial Networks (GAN) are usually used to enhance data [24] to make up for the difference in the number of data types. However, these enhancement methods improve the efficiency of model training by increasing the number of samples, and do not consider the color differences in histopathological images. This article takes another approach, using the mixed data method to help the model learn high-level semantic features in pathological images by generating a pretext task, so this article is also a new attempt in the field of pathological data enhancement.

2.3 SSL for medical image

Due to the particularity of medical image data and the fact that it is much easier to collect large-scale unlabeled medical image datasets than hand-labeled small datasets, there are mainly three types of SSL methods: predictive SSL, generative SSL and comparative study.

For predictive SSL, Lu et al. [25] used contrastive predictive coding and multi-instance learning to classify breast cancer histological images, where contrastive predictive coding was used to learn rich representations from breast cancer histopathological images; Generative SSL, Hervella et al. [26] propose a new alternative that allows the application of transfer learning from unlabeled data in the same domain, including using multimodal reconstruction tasks. Experimental results show that the Self-Supervised transfer learning strategy achieves state-of-the-art (SOTA) performance in all research tasks; for contrastive learning, Yang et al. [27] employed two proxy tasks for SSL, namely generative cross-coloring prediction and discriminative contrastive learning. They both leverage domain-specific knowledge well and do not require side information. Yan et al. [28] present a robust and label-efficient self-supervised FL framework for medical image analysis. Experimental results show that masked image modeling with Transformers significantly improves the robustness of models against various degrees of data heterogeneity. Taleb et al. [29] propose the ContIG, a self-supervised method that can learn from large datasets of unlabeled medical images and genetic data. The results show that including genetic information in the pre-training process can significantly improve performance.

3. METHODOLOGY AND MATERIAL

3.1 MAE

Histopathological imagery, distinct from natural scenes, encapsulates hierarchical information across various magnification levels, embodying specific structures and features at both macroscopic and cellular dimensions. These images harbor an elevated tier of information, predominantly concerning pathologies and anomalies.

The MAE facilitates the reconstruction of such images through strategic pixel omission, compelling the neural architecture to assimilate a deeper understanding of the image's essence. The deployment of masks is pivotal, steering the focus of reconstruction. A sparing use of masks hones in on the restoration of intricate details, akin to super-resolution endeavors, whereas a generous application of masks aims to unearth global semantic insights.

In the context of refurbishing pathological images with MAE, it's imperative to opt for a more substantial mask ratio. This approach nudges the encoder towards recognizing and internalizing high-level semantic content, a crucial advantage for subsequent analytical and diagnostic procedures.

3.2 Strategies before the mask

As delineated in Figure 2, our investigation harnesses H&E-stained histopathological images of lung cancer sourced from a plethora of medical institutions and specimens. Nonetheless, divergences in tissue sectioning techniques and imaging modalities may engender disparate color representations across these images. Furthermore, the digital conversion process of these specimens accentuates color variances, attributable to a multitude of factors including sample illumination, magnification levels, image capturing nuances, compression algorithms, storage conditions, and display technologies.

To mitigate the pronounced disparities in the visual presentation of digital pathological slices, we employ the Mixup technique to diminish the color differentiation's influence on the model, thereby bolstering its resilience. Mixup is a data augmentation technique that creates synthetic images by combining pairs of original images through weighted blending of their pixel values. This method not only increases the diversity of the training data but also enables the network to better handle the complex and varying colors in the dataset.

Additionally, the synthetic generation of pathological tissue images via Mixup empowers the network to proficiently extract high-level semantic information from the data's complexity. This strategy fosters the model's ability to assimilate knowledge from a broader and more heterogeneous collection of pathologies, potentially augmenting its efficacy in identifying lung cancer.

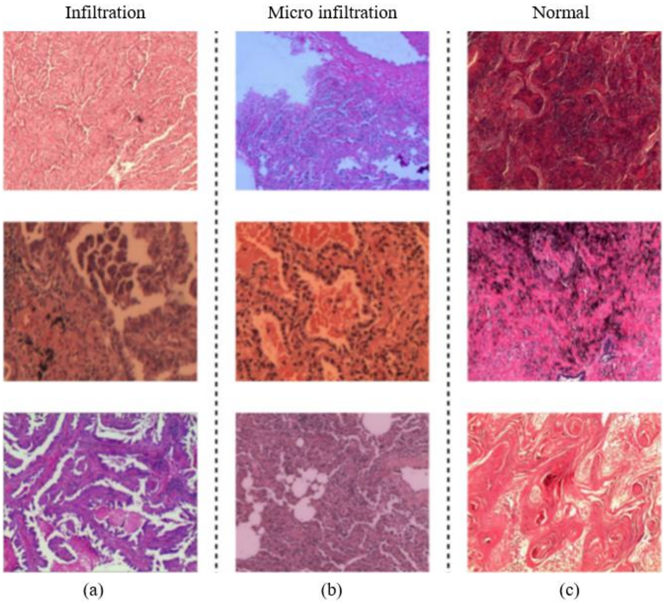


Figure 2. Differences in the appearance of digital pathology slides: (a) Pathological picture of infiltration lung cancer; (b) Pathological picture of micro infiltration lung cancer; (c) Normal lung samples

Employing the Mixup methodology, we perform a pixel-level weighted amalgamation of randomly selected pathological images, incorporating these augmented, mixed-sample data as virtual samples for model training. For two training sample images and randomly selected from the training set and we have:

$$x' = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

where, $\lambda \in [0,1]$ represents the weight of the sample. x' represents the result of a blend operation of two training sample images. The inter-sample region performed by the linear weighting process by Mixup enables the model to learn additional samples besides the training samples, thus reducing the inadaptability of data prediction targets beyond the training samples and providing smoother uncertainty estimates. As shown in Figure 3, although it constructs a virtual pathological image that partly does not exist in reality, it is observed from the mixed image that it can indeed mix different features in the pathological image by weight.

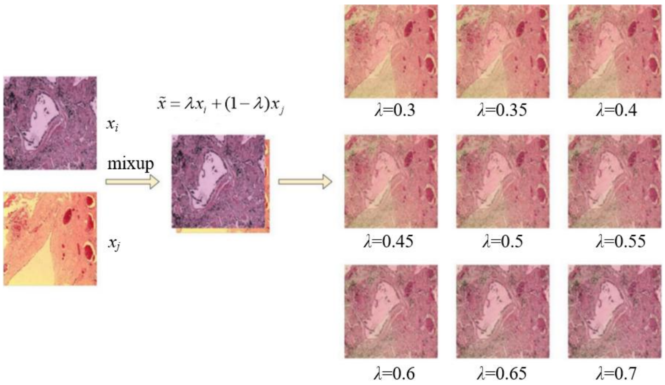


Figure 3. Schematic diagram of pathological image Mixup

Compared with traditional strategies such as MAE masking and Mixup without mixing objectives, the MixMAE method can learn more features. Furthermore, compared to models trained without data augmentation or with different augmentation, models trained with Mixup show higher stability, which can be used to address overfitting in medical pathology images.

3.3 Overview of MixMAE

Traditional supervised learning requires large amounts of labeled data, which in many cases is expensive and time-consuming to obtain. The performance of a model is highly dependent on the quality of the labeled data. If the data is inaccurately or inconsistently labeled, the performance of the model may be severely affected. The model may overfit the training data, resulting in a reduced ability to generalize over unseen data. For a SSL model, on the other hand, the ability to learn from unlabeled data means that it can utilize a large amount of existing data without spending a lot of time and resources on data labeling. SSL models typically have better generalization capabilities because they are trained on a wider distribution of data rather than being limited to specific labeled datasets. In addition, with SSL, models can learn deep features and structures of the data that are useful for subsequent tasks (e.g., classification, detection, etc.). Therefore, we built a SSL model for performing a downstream lung cancer pathology image classification task.

MixMAE introduces a cutting-edge hybrid Self-Supervised visual representation learning framework, illustrated in Figure 1. This framework unfolds across two critical stages: an initial Self-Supervised pre-training phase targeting upstream tasks, followed by a supervised fine-tuning phase for downstream tasks. To enhance the network's generalization and robustness, it begins by blending original H&E stained histopathological images with two distinct pathological images for data augmentation. In the first stage, this augmented dataset is partially masked, and the MAE is tasked with a novel generative proxy task: reconstructing the obscured pixels, thereby predicting histopathological images. The encoder within MAE utilizes a Transformer architecture to efficiently learn latent features from the histopathological images. Progressing to the second stage, the approach is inspired by the ViT, repurposing the encoder refined during pre-training alongside a MLP tailored for the classification task. This phase involves fine-tuning the model with a limited set of labeled data, striking a balance between SSL's broad applicative scope and supervised learning's precision. The core equation of the MixMAE model is as follows:

1. Data enhancement phase: The original images (x_i) and (x_j) are mixed by Mixup technique to generate the enhanced image (x') from the Eq. (1).

2. Self-supervised pre-training phase: Randomize the mask on the enhanced image (x') to generate a masked image (o). MAE uses the Transformer encoder (E) to learn the latent feature representation (z) from the masked image (o):

$$z=E(o) \tag{2}$$

3. Supervisory fine-tuning phase of downstream missions: In the subsequent downstream task of classifying different types of lung cancer pathology images, the labeled data were fine-tuned using the pre-trained encoder (E) and MLP classifier (C) to obtain the predicted labels (Y):

$$Y=C(E(x)) \tag{3}$$

The ViT is regarded as the encoder backbone structure. The network consists of a self-attention mechanism and a Multi-Layer Perceptron (MLP), which can properly model the global relations. The input of the pathological image is $x \in \mathbb{R}^{C\times H\times W}$, where C represents the Channel of the input image, and H and W represent the length and width of the image, respectively. The image of $H\times W$ pixels is divided into $\frac{H\times W}{P^2}$ regular non-overlapping patches by Embedding Patches, where P represents the size of the patch. Then through uniform random sampling to cover some patches, the image features of the visible part are embedded by linear projection, and the embedded features and position information (W^Q, W^K, W^V) are converted into Queries ($Q=W^Q\times x$) and Keys ($K=W^K\times x$) and Values ($V=W^V\times x$), then Q, K, and V are fed into the Transformer block. Self-attention is the core of the Transformer block, and the expression is as follows:

$$\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Firstly, the Query-Key pair is used by self-attention to measure the attention matrix of the visible patch, and the attention weight is obtained through Softmax. Finally, the value is weighted and summed according to the weight coefficient to calculate the self-attention of the image to extract the potential representation of the visible part.

During pre-training, the MAE decoder is tasked with image reconstruction, while the encoder focuses on generating image representations for recognition tasks. This allows for a flexible decoder design, independent of the encoder's architecture. Our experiments reveal a smaller, less complex decoder is effective, being narrower and shallower than the encoder.

In the pre-training phase, the model engages with images altered by Mixup, blending pixels from different images. This approach enables the model to extract more information from the visible segments of these blended images. The pixel-mixed images possess RGB values distinctly different from the original, diminishing color-related discrepancies and facilitating the model's performance in downstream classification tasks. This method enhances the model's ability to learn from limited data while reducing color bias, ensuring robustness and accuracy in classification.

3.4 Data augmentation

In order to be able to flexibly cope with the complexity of clinical processing, we used data augmentation techniques to enhance the heterogeneity of the relevant images in the lung cancer pathology training set. Data enhancement can increase the number and diversity of pathology samples in the training set, improve the stability of model performance under noisy data, and improve the generalization ability of neural networks.

Data augmentation diversifies the training dataset through techniques such as random horizontal and vertical flips, rotations, and noise addition. Each augmented image undergoes meticulous inspection to confirm the presence of regions pertinent to lung cancer pathology.

This strategy of integrating random variations fortifies the model against noisy data, fostering improved generalization and performance. To enrich the dataset's diversity and volume, images undergo five distinct augmentations, broadening the model's exposure to various pathological features. For consistency, all enhanced samples are resized to 224×224, optimizing the training process.

3.5 LC25000

The public dataset of histopathological images of lung and colon cancer (LC25000) contains 25000 color images, which are divided into five categories, namely colon adenocarcinoma, benign colon tissue, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue, 5000 images per class [30]. All images are 768×768 pixels in size. Figure 4 shows the representative histopathological images of the three categories respectively, and Table 1 shows the images of each category are divided into training set and test set according to the ratio of 4:1. A total of 25000 histopathological images, 20000 for training and 5000 for testing.

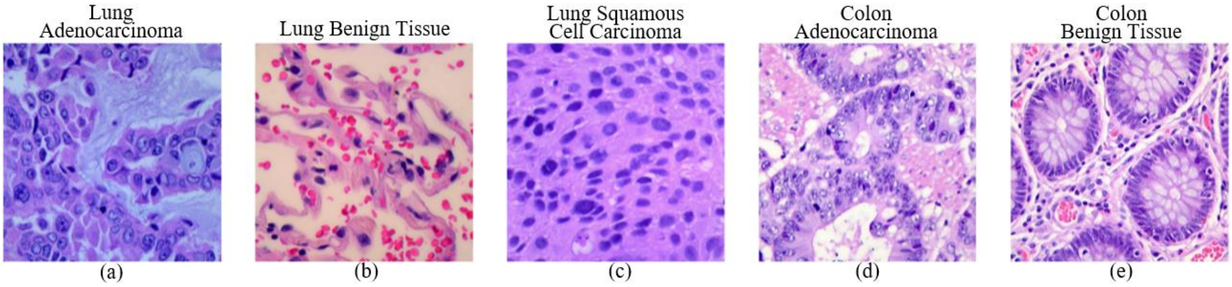


Figure 4. LC25000 public dataset: (a) Lung Adenocarcinoma (Lung_aca); (b) Lung Benign Lung (Lung_n); (c) Lung Squamous Cell Carcinoma (Lung_scc); (d) Colon Adenocarcinoma (Colon_aca); (e) Colon Benign Tissue (Colon_n)

Table 1. Data distribution of five types of samples in LC25000

Image Type	Train	Test	Sum
Lung_aca	4000	1000	5000
Lung_n	4000	1000	5000
Lung_scc	4000	1000	5000
Colon_aca	4000	1000	5000
Colon_n	4000	1000	5000

3.6 Lung

The Bethune First Hospital of Jilin University contributed lung cancer pathological digital slide image data from 760 cases over 2021-2022. This dataset was meticulously curated and annotated over five months by three experts, each boasting over five years of professional experience, ensuring the clinical nuances of each sample were pronounced. Digital pathological scanning equipment transformed the selected pathological specimens into digital images, which were then segmented into slices at 20× magnification, each measuring 2048×1500 pixels. During data preprocessing, slices exhibiting blurring, overstaining, or undesirable slide backgrounds were systematically excluded through manual inspection.

Furthermore, the dataset was enriched through data augmentation techniques, including random rotations, flips, and selective occlusion, to enhance the diversity of the training set. Consequently, as delineated in Table 2, a total of 7,932 lung cancer pathological slice images met the criteria for inclusion in this study, stratified into 2,644 infiltration cases, 2,644 micro-infiltration diagnoses, and 2,644 instances classified as normal tissue. From this corpus, 870 images (comprising 290 from each category) were randomly selected to constitute the test set.

Table 2. Data distribution of five types of samples in lung

Image Type	Train	Test	Sum
Infiltration	2354	290	2644
Micro Infiltration	2354	290	2644
Normal	2354	290	2644

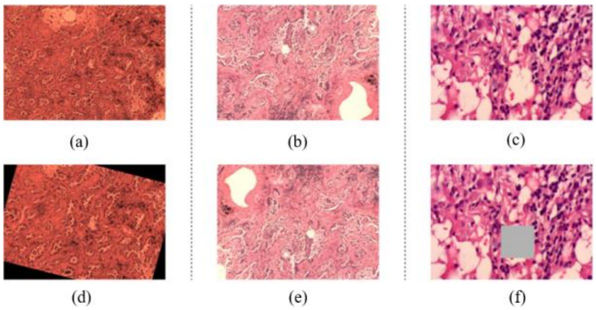


Figure 5. Private data of lung adenocarcinoma: (a) Infiltration Lung Adenocarcinoma; (b) Micro Infiltration Lung Adenocarcinoma; (c) Normal Lung Tissue; (d) Random sample rotation; (e) Random flip; (f) Random area occlusion

Figure 5 illustrates the distinct tumor growth patterns characteristic of infiltration and micro-infiltration, where the former adheres and expands beyond a 0.5 cm diameter, displaying acinar, papillary, micropapillary, or solid structures. Conversely, micro-infiltration lesions remain confined within a 0.5 cm diameter. The term 'normal' is used to denote benign regions within lung tissue, underscoring the dataset's comprehensive scope in capturing the spectrum of lung cancer pathology.

4. RESULT

4.1 Experiment settings

To evaluate the performance and effectiveness of the MixMAE model on lung cancer pathological images, we conducted a series of experiments. The experimental environment and hyperparameters were appropriately configured. For the hyperparameters in both the upstream pre-training tasks and downstream model fine-tuning tasks, we used the AdamW optimizer to adjust the network parameters. The batch size was set to 512, and the initial learning rate was set to 0.0001 (decaying 10 times every 20 steps). Additionally, the upstream pre-training task was completed within 600 epochs, and the downstream model fine-tuning task was completed within 200 epochs.

Both models were developed within the PyTorch 1.8.0 framework, leveraging NVIDIA CUDA v8.0 and cuDNN v10.1 libraries for acceleration, and coded in Python 3.7. These experiments were conducted on a Windows 10 platform, powered by an Intel Core i9-10875H CPU at 2.30 GHz, an NVIDIA RTX 3090 GPU, and 32 GB of RAM, ensuring optimal computational efficiency and reliability in processing.

4.2 Evaluation index

To evaluate the diagnostic performance of the model for cancer pathology, we used the overall Accuracy, Precision, Sensitivity, and Specificity as the evaluation metrics and compared the results of the proposed method with those of the state-of-the-art model. The overall accuracy indicates the proportion of samples correctly predicted by the model to the total number of samples, the precision indicates the proportion of samples correctly predicted by the model to the total number of positive samples, the sensitivity indicates the proportion of samples correctly predicted by the model to the total number of positive samples, and the specificity indicates the proportion of samples correctly predicted by the model to the total number of negative samples:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{7}$$

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

Among them, TP (True Positive) indicates that an instance is a positive sample and is also predicted as a positive sample; FN (False Negative) indicates that an instance was originally a positive sample but was predicted as a negative sample; FP (False Positive) indicates that an instance was originally a negative sample, but it is judged as a positive sample; TN (True Negative) indicates that an instance is a negative sample and is also judged as a negative sample.

4.3 Experimental results on private datasets

In addressing the classification of lung cancer pathological images, we evaluated our MixMAE model alongside BEiT [30], MoCov3 [31], and MAE, with the latter three serving as archetypes of SSL models. MixMAE, also grounded in SSL, was compared within the same framework, which bifurcates into upstream and downstream tasks. The upstream task involves feature extraction from the unlabeled pathological image dataset, partitioning images into blocks for batch processing over numerous iterations, enabling the model to identify pivotal features. The downstream task leverages these acquired features to categorize test set images, fulfilling the classification objective. Uniform training hyperparameters were applied across all models, with comparative outcomes presented in Table 3. These data showed that the MixMAE model performed best on all four assessment metrics, especially on Specificity, which reached the highest 97.33%. In contrast, the BEiT model performed relatively low on these four metrics, with Accuracy, Precision, Sensitivity, and Specificity of 90.81%, 91.10%, 90.80%, and 95.40%, respectively. This indicates that the MixMAE model has high accuracy in distinguishing negative samples in lung adenocarcinoma pathology images.

Table 3. Test results of four self-supervised networks on private datasets

Network	Accuracy	Precision	Sensitivity	Specificity
BEiT [31]	90.81	91.10	90.80	95.40
MoCov3 [32]	92.60	92.52	93.04	96.15
MAE [12]	94.52	94.23	93.31	96.88
MixMAE	95.64	94.73	93.88	97.33

It can be found through Figure 6 that the accuracy of both MAE and MixMAE starts to stabilize around 170epoch. Compared with the original MAE, MixMAE enriches the image information and improves the accuracy from 94.52 to 95.64, precision from 94.23 to 94.73, sensitivity by 0.57, and specificity by 0.45. In addition, without increasing the number of parameters, MixMAE can obtain better results than MAE. Compared with BEiT and MoCov3, MixMAE is more effective and the improvement is more obvious.

The experimental results can be further analyzed through the related model confusions. As can be seen in Figure 7, the errors in the classification process of the four models are mainly concentrated in the confusion between immersion and micro-immersion. The amount of data for which the four models, BEiT, MoCov3, MAE, and MixMAE, make incorrect predictions for immersion and micro-immersion data are 48, 38, 33, and 28, respectively. The small difference between the two types of samples leads to the fact that the models provide an incorrect judgments, which intuitively is consistent with our observed error characteristics, and the training accuracy of the normal sample model is very high. Conversely, it can also be seen from the figure that MixMAE discriminates infiltration and microinfiltration to a higher degree than the other models, with fewer erroneous judgments in the infiltration and microinfiltration categories, due to the fact that MixMAE itself learns better than the other models in the upstream task. Its learning ability is stronger than other models, and thus it has the highest accuracy in the downstream task. Therefore, the different pathological features of infiltration and microinfiltration can be better recognized using MixMAE.

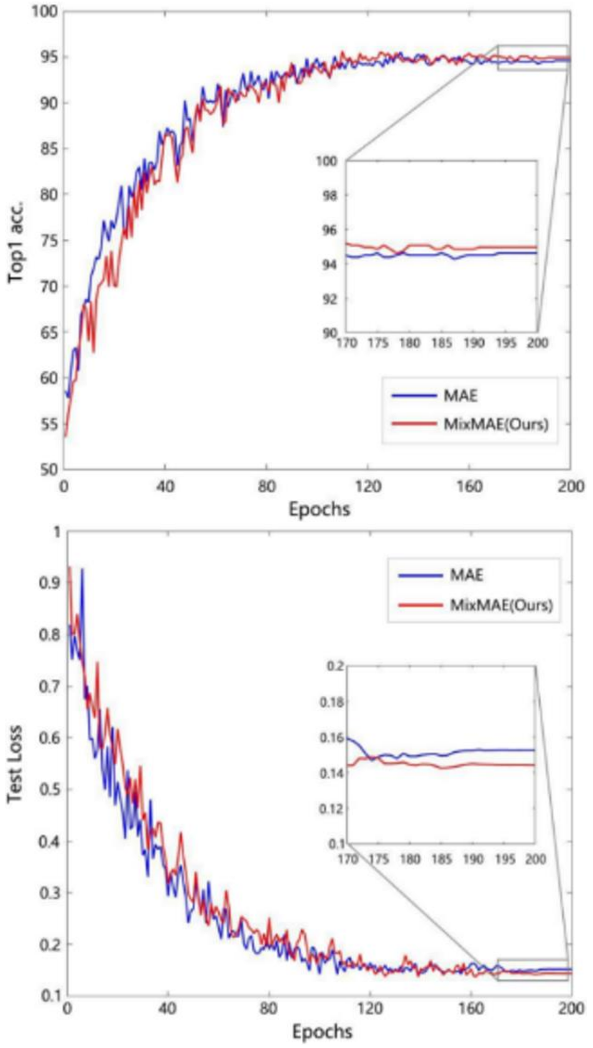
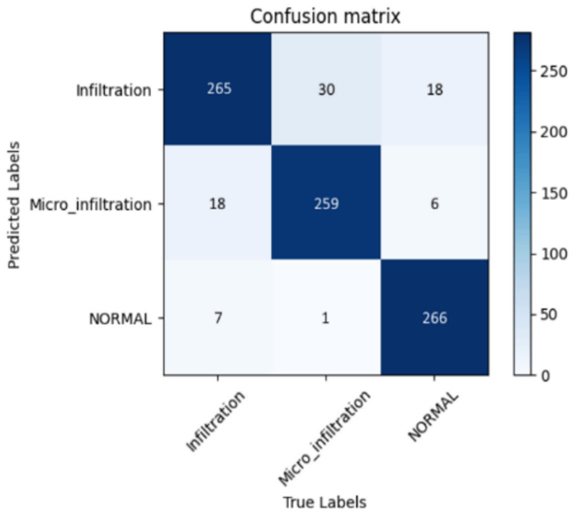
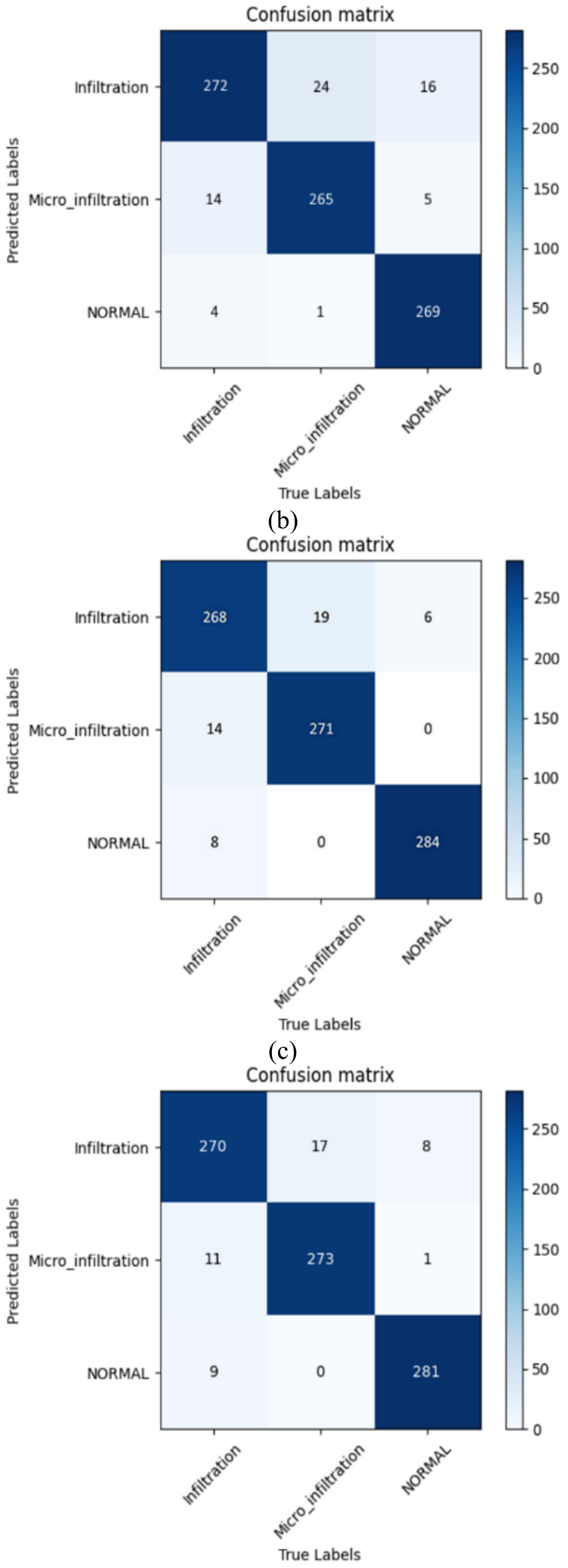


Figure 6. Comparison of the accuracy and loss rate of MAE and MixMAE in the downstream 200 rounds of testing



(a)



(b)

(c)

(d)

Figure 7. The confusion matrix obtained by the model on a private dataset: (a) BEiT; (b) MoCov3; (c) MAE; (d) MixMAE

4.4 Ablation experiment

The above results show that MixMAE can better diagnose invasive and microinvasive lung cancer due to its good feature learning ability. Therefore, we conduct ablation experiments by changing the size of the mask rate to observe the effect of the mask rate on the model. All experiments were performed with 600 epochs of upstream training and 200 epochs of downstream testing. While keeping the experimental parameters uniform, only the masking ratios were changed to 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. The evaluation indicators are accuracy, upstream loss rate (L_{mse}) and downstream loss rate (L_{cls}). The comparison results are shown in Table 4.

Table 4. MixMAE ablation experiments with mask rates ranging from 0.4 to 0.9.

Mask Ratio	Accuracy	L_{mse}	L_{cls}
0.4	95.18	0.831	0.150
0.5	95.52	0.821	0.134
0.6	95.41	0.819	0.142
0.7	95.64	0.815	0.137
0.8	95.59	0.818	0.142
0.9	94.97	0.826	0.153

The outcomes of our ablation studies illuminate that the model attains peak accuracy and learning efficiency when the mask rate oscillates between 0.7 and 0.8. This observation underscores the principle that veiling a more significant portion of the input images engenders a more profound learning impact. Conversely, should the mask rate eclipse 0.8, a decline in model performance becomes evident. Similarly, a mask rate beneath 0.4 also precipitates suboptimal outcomes, attributed to the insufficiency of masks available for the model's learning process.

This pattern holds true for the upstream loss rate, mirroring the trends observed in accuracy. The most favorable upstream loss rate coincides with a mask rate nestled within the 0.7 to 0.8 range, signifying this interval as the most conducive for model training. Deviations beyond or below this range are synonymous with heightened loss rates.

Notably, a mask rate of 0.5 emerged as the dark horse, revealing the lowest training loss and an augmented accuracy for downstream tasks. This suggests that a mask rate of 0.5, balancing between excessive and inadequate masking, may better serve image reconstruction endeavors, bolstering the model's proficiency in regenerating the obscured segments of the images.

Conclusively, these insights collectively advocate for a mask rate corridor of 0.7 to 0.8 as the zenith for optimizing the MixMAE model in the task of classifying lung cancer pathological images, marking a fine line between too much and too little, where the model's learning and predictive capabilities are maximally harnessed.

4.5 Extended experiment

To validate the MixMAE model's training impartiality beyond lung cancer pathological images, its generalization prowess was assessed using the LC25000 dataset, encompassing both lung and colon cancer pathological images across five classifications: conventional lung digital pathology images (Lung_n), lung adenocarcinoma (Lung_aca), lung squamous adenocarcinoma (Lung_scc), colon adenocarcinoma (Colon_aca), and normal colon cells (Colon_n). Each category boasts 5,000 samples, cumulating in a comprehensive tally of 25,000 samples, underpinning the experimental findings' reliability.

Echoing the methodology applied to the proprietary lung cancer dataset, the LC25000 was partitioned into training and testing subsets, consisting of 4,000 and 1,000 images per category, respectively. Each sample within the LC25000 dataset was meticulously annotated. To ensure experimental equity, the model underwent training on this public dataset under identical parameters as those utilized for the private dataset.

The confusion matrix shown in Figure 8 illustrates the results achieved by MixMAE in the classification task on the public dataset LC25000. It can be seen that MixMAE performs very well with an accuracy close to 100%. Only 2 pathology images out of 5000 data for three lung pathology images and two colon pathology images were incorrectly predicted. This shows that the MixMAE model has a strong recognition ability for both lung and colon cancers, verifies that the MixMAE model has a good generalization ability, and lays the foundation for MixMAE to be applied in the diagnostic task of other cancer pathology images.

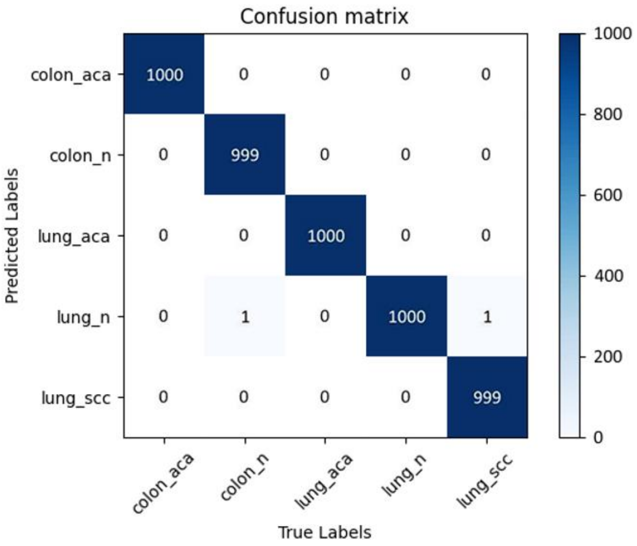


Figure 8. Extended experiment of MixMAE on public dataset LC25000

5. DISCUSSION

The cornerstone of this research lies in the innovative fusion of Mixup and MAE—two potent techniques for data augmentation and feature extraction—culminating in the pioneering SSL framework, MixMAE, tailored for lung cancer pathology image analysis. This framework stands out from prior approaches with several distinct advantages:

1. Mixup Method: By employing the Mixup method to blend diverse pathological images, this study enriches the dataset's diversity, mitigates the influence of color variations, and bolsters the model's stability and resistance to interference. In contrast, conventional studies often rely on single-image inputs, neglecting the wealth of latent data information, which can lead to overfitting and compromised generalization capabilities.

2. MAE Method: The MAE method's strategic pixel masking compels the model to reconstruct occluded segments based on visible regions, thereby enhancing its grasp of both global and local image attributes and bolstering its expressive and reconstructive prowess. Traditional studies typically harness convolutional neural networks for feature extraction, failing to fully exploit the structural and semantic nuances of images, which can result in deficient expressive and reconstructive faculties.

3. ViT Backbone: Integrating the Vision Transformer (ViT)

as the backbone within the MAE encoder leverages the self-attention mechanism and multilayer perceptron to adeptly capture long-range image dependencies and extract profound image features. ViT's computational efficiency and reduced parameter count, compared to traditional convolutional networks, allow for greater adaptability to varying image sizes and resolutions, thus enhancing the model's flexibility. Previous studies often fixate on static image sizes and resolutions, overlooking the potential of scaling and detail, which can impede the model's adaptability.

Despite these advancements, the study acknowledges certain limitations that pave the way for future enhancements. The pixel-level mixing of the Mixup method might obscure fine details, impacting the model's reconstruction fidelity and recognition accuracy. Subsequent research could delve into feature-level or semantic-level mixing to preserve more salient information, thereby refining the model's expressiveness and interpretability. Additionally, the random masking inherent in the MAE method might overlook critical regions or features, affecting the learning outcomes and generalization. Future endeavors might investigate attention-driven or saliency-based masking to direct the model's focus toward more significant elements, thus improving learning efficiency and generalization.

Extending MixMAE to other cancer pathologies, such as breast, liver, or stomach cancer, could validate its universal applicability and adaptability across various contexts, offering a comprehensive framework for cancer pathology image analysis. Moreover, integrating MixMAE with other SSL paradigms—like contrastive, clustering, or generative learning—could broaden the horizons of SSL objectives and loss functions, further advancing the model's self-learning and representational capabilities.

6. CONCLUSIONS

In this study, we present MixMAE, an innovative SSL framework that synergistically integrates Mixup image augmentation with MAE. This approach is particularly adept at harnessing the limited labeled data available in pathological image analysis, extracting advanced semantic information through a novel pretext task. The salient contributions of our work are:

1. Innovative Framework: MixMAE stands as a pioneering algorithm within the realm of SSL, merging the strengths of MAE's feature extraction with Mixup's data augmentation. This fusion facilitates the learning of a richer feature set, enhancing the model's performance in subsequent classification tasks.

2. Enhanced Feature Complexity: The features discerned by MixMAE surpass those identified by MAE in both quantity and complexity. This enables MixMAE to adeptly handle intricate images and excel in challenging datasets.

3. Clinical Relevance: Utilizing a curated dataset of lung cancer infiltration and micro-infiltration, MixMAE has demonstrated its prowess in accurately distinguishing between these nuanced features, underscoring its potential utility in real-world clinical scenarios.

To encapsulate, MixMAE embodies a significant leap forward in SSL for pathological image analysis. It not only elevates the model's efficiency without additional computational burdens but also achieves superior accuracy and diagnostic metrics. Our extensive validation on diverse datasets, coupled with rigorous ablation studies, confirms MixMAE's superiority over existing methods. Its adeptness at differentiating critical cancer features heralds a new horizon for clinical diagnostics, paving the way for its deployment in practical healthcare settings and inspiring future research directions in medical image analysis.

REFERENCES

[1] Siegel, R.L., Miller, K.D., Goding Sauer, A., et al. (2020). Colorectal cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(3): 145-164. https://doi.org/10.3322/caac.21601

[2] Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B. (2009). Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering, 2: 147-171. https://doi.org/10.1109/RBME.2009.2034865

[3] Jara-Lazaro, A.R., Thamboo, T.P., Teh, M., Tan, P.H. (2010). Digital pathology: Exploring its applications in diagnostic surgical pathology practice. Pathology, 42(6): 512-518. https://doi.org/10.3109/00313025.2010.508787

[4] Madabhushi, A., Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis, 33: 170-175. https://doi.org/10.1016/j.media.2016.06.037

[5] Rashidi, H.H., Tran, N.K., Betts, E.V., Howell, L.P., Green, R. (2019). Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. Academic Pathology, 6: 2374289519873088. https://doi.org/10.1177/2374289519873088

[6] Lee, Y., Park, J.H., Oh, S., et al. (2022). Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nature Biomedical Engineering. https://doi.org/10.1038/s41551-022-00923-0

[7] Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. Scientific Reports, 7(1): 4172. https://doi.org/10.1038/s41598-017-04075-z

[8] Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine, 27: 317-328. https://doi.org/10.1016/j.ebiom.2017.12.026

[9] Liu, Y., Wang, H., Song, K., et al. (2022). CroReLU: cross-crossing space-based visual activation function for lung cancer pathology image recognition. Cancers, 14(21): 5181. https://doi.org/10.3390/cancers14215181

[10] Yucel, N., Mutlu, H.B., Durmaz, F., Cengil, E., Yildirim, M. (2023). A CNN approach for enhanced epileptic seizure detection through EEG analysis. Healthcraft Frontiers, 1(1): 33-43. https://doi.org/10.56578/hf010103

[11] Chen, H., Li, C., Wang, G., et al. (2022). GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recognition, 130: 108827. https://doi.org/10.1016/j.patcog.2022.108827

[12] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.

(2022). Masked autoencoders are scalable vision learners. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 15979-15988. https://doi.org/10.1109/CVPR52688.2022.01553

[13] Azizi, S., Mustafa, B., Ryan, F., et al. (2021). Big self-supervised models advance medical image classification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 3458-3468. https://doi.org/10.1109/ICCV48922.2021.00346

[14] Li, X., Hu, X., Qi, X., Yu, L., Zhao, W., Heng, P.A., Xing, L. (2021). Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. IEEE Transactions on Medical Imaging, 40(9): 2284-2294. https://doi.org/10.1109/TMI.2021.3075244

[15] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis, 58: 101539. https://doi.org/10.1016/j.media.2019.101539

[16] Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems, 33: 12546-12558.

[17] Ciga, O., Xu, T., Martel, A.L. (2022). Self supervised contrastive learning for digital histopathology. Machine Learning with Applications, 7: 100198. https://doi.org/10.1016/j.mlwa.2021.100198

[18] Zhang, Y., Ye, Y., Wang, D. (2003). Clinical application of image processing and neural network in cytopathological diagnosis of lung cancer. Chinese Journal of Thoracic and Cardiovascular Surgery, 2003(14): wpr-573919.

[19] Yang, H., Chen, L., Cheng, Z., et al. (2021). Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. BMC medicine, 19: 80. https://doi.org/10.1186/s12916-021-01953-2

[20] Kanavati, F., Toyokawa, G., Momosaki, S., et al. (2020). Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific Reports, 10(1): 9297. https://doi.org/10.1038/s41598-020-66333-x

[21] Chen, Y., Feng, J., Liu, J., Pang, B., Cao, D., Li, C. (2022). Detection and classification of Lung Cancer cells using swin transformer. Journal of Cancer Therapy, 13(7): 464-475. https://doi.org/10.4236/jct.2022.137041

[22] Wang, J., Liu, X. (2021). Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network. Computer Methods and Programs in Biomedicine, 207: 106210. https://doi.org/10.1016/j.cmpb.2021.106210

[23] Chen, H., Li, C., Li, X., et al. (2022). IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. Computers in Biology and Medicine, 143: 105265. https://doi.org/10.1016/j.compbiomed.2022.105265

[24] Saha, M., Guo, X., Sharma, A. (2021). Tilgan: Gan for facilitating tumor-infiltrating lymphocyte pathology image synthesis with improved image classification. IEEE Access, 9: 79829-79840. https://doi.org/10.1109/ACCESS.2021.3084597

[25] Lu, M.Y., Chen, R. J., Mahmood, F. (2020). Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation). Medical Imaging 2020: Digital Pathology, 11320: 113200J. https://doi.org/10.1117/12.2549627

[26] Hervella, Á.S., Rouco, J., Novo, J., Ortega, M. (2020). Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. Applied Soft Computing, 91: 106210. https://doi.org/10.1016/j.asoc.2020.106210

[27] Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., Lv, H. (2022). CS-CO: A hybrid self-supervised visual representation learning method for H&E-stained histopathological images. Medical Image Analysis, 81: 102539. https://doi.org/10.1016/j.media.2022.102539

[28] Yan, R., Qu, L., Wei, Q., et al. (2023). Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. IEEE Transactions on Medical Imaging, 42(7): 1932-1943. https://doi.org/10.1109/TMI.2022.3233574

[29] Taleb, A., Kirchler, M., Monti, R., Lippert, C. (2022). Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 20876-20889. https://doi.org/10.1109/CVPR52688.2022.02024

[30] Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M. (2019). Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142. https://doi.org/10.48550/arXiv.1912.12142

[31] Bao, H., Dong, L., Piao, S., Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254. https://doi.org/10.48550/arXiv.2106.08254

[32] Chen, X., Xie, S., He, K. (2021). An empirical study of training self-supervised vision transformers. https://doi.org/10.48550/arXiv.2104.02057