Weakly supervised contrastive learning infers molecular subtypes and recurrence risk of breast cancer from pathology images

Hui Liu[1], Yang Zhang[1] and Judong Luo[2*]

[1]School of Computer Science and Technology, Nanjing Tech University, Nanjing, 211816, Jiangsu, China.
[2]Department of Radiotherapy, The Affiliated Changzhou Second People's Hospital of Nanjing Medical University, City, 213110, Jiangsu, China.

*Corresponding author(s). E-mail(s): judongluo@163.com;

Abstract

The integration of deep learning with ultra-high resolution digital pathology has led to significant improvements in the speed and accuracy of tumor diagnosis, while also demonstrating substantial potential to infer genetic mutations and gene expression levels. However, the application of computational pathology in predicting molecular subtypes, prognosis, and recurrence risk remains limited in breast cancer. To address this issue, our paper proposes a weakly supervised contrastive learning method. This method first performs contrastive learning pre-training on a large-scale unlabeled pathological image to extract tile-level features. The tile-level features are then aggregated using a gated attention mechanism to obtain slide-level features. During the downstream tasks, the feature extractor is fine-tuned using weak supervisory signals. To confirm the effectiveness of this method, we conducted extensive experiments, including tumor diagnosis, gene expression level prediction, molecular typing, recurrence, and drug response prediction, as well as a prognostic risk score. Our findings indicate that our method can accurately identify tumor regions using only weakly supervisory signals and performs exceptionally well in the prediction of breast cancer-related gene expression levels, as well as multiple clinically relevant tasks. Additionally, the tile-level attention scores learned in downstream tasks are used to visualize spatial heat maps that are highly consistent with pathologist annotations and spatial transcriptome.

These results demonstrate that the deep learning-based model effectively establishes the high-order genotype-phenotype associations, thereby enhances the potential of digital pathology in clinical applications.

Keywords: Contrastive learning, pathological imaging, RNA-seq, weakly supervised learning, molecular subtyping, tumor recurrence

1 Introduction

Breast cancer is one of the most prevalent and lethal malignant neoplasms among women. According to recent statistics, in 2020 there were 2.6 million cases of breast cancer [1]. Breast cancer exhibits considerable tumor heterogeneity, with marked variability in molecular subtypes, treatment outcomes, and prognosis among individual patients. At present, genetic testing is the standard methodology for molecular subtyping and pre-treatment efficacy assessment. However, the high cost and long turnaround time of genetic testing impede its widespread application. Consequently, there is an pressing demand in clinical practice for the development of a new, cost-effective and scalable alternative approach.

With the advent of whole-slide imaging technology and the growing emphasis on precision medicine, AI-powered digital pathology has facilitated the extraction of histological features beyond human visual perception ability. Digital pathology has emerged as an efficacious approach to inferring genetic alteration and molecular signatures. For example, Noorbakhsh et al. developed a convolutional neural network (CNN) model that effectively differentiated between tumor/normal tissue and cancer subtypes among 19 types of cancer[2]. Wang et al. proposed a deep learning model to study the tumor microenvironment using pathological images and uncovered that cellular spatial organization is predictive of gene expression profiles and patient survival[3]. Wang et al. developed a deep learning-based histological grading model (DeepGrade) based on whole-slide images (WSI) to improve risk stratification for NHG 2 breast cancer patients and validated the effectiveness of their model [4]. Kather et al. employed a deep residual network to directly predict microsatellite instability from pathological images[5]. Jain et al. utilized deep learning to predict tumor mutation burden from pathological images [6]. Saltz et al. leveraged deep learning to reveal the tumor immune microenvironment and lymphocyte infiltration[7].

The aforementioned studies illustrate the remarkable potential of deep learning in the realm of digital pathology image analysis. However, this approach is also confronted with the challenge of a paucity of pixel-level manual annotation, as it is both costly and time-consuming, due to experienced pathologists are exceedingly scarce and busy. To address this challenge, some researchers have turned to semi-supervised and weakly supervised learning techniques. For example, Mahmood et al. developed the CLAM model,

which utilizes weakly supervised learning and attention mechanisms to classify subtypes of renal cell carcinoma and non-small cell lung cancer using pathological images[8]. Zhang et al. devised a dual-layer feature distillation multi-instance learning model that demonstrated impressive performance on the CAMELYON-16 and TCGA lung cancer datasets[9]. Wu et al. harnessed the power of consistency regularization and entropy minimization to develop an efficacious semi-supervised kernel image segmentation algorithm[10].

In recent years, the rapid advancement of multi-omics technology has provided novel insights into the tumor molecular subtypes, biomarkers of prognosis and drug response. For instance, gene expression profiling has facilitated the establishment of molecular subtypes of breast cancer[11]. A 12-gene expression signature has been developed as an independent indicator of recurrence for stage II colon cancer patients [12]. Furthermore, gene signatures have been employed to establish recurrence risk models for early-stage breast cancer patients [13][14][15]. However, due to its high cost, molecular diagnosis based on gene sequencing cannot be applied widespread in clinical practice. Consequently, researchers have endeavored to infer gene expression profiles based on high-resolution pathological images. For example, Kather et al. demonstrated that pan-cancer genetic alterations could be inferred from pathological images [16]. Schmauch et al. proposed the HE2RNA model to predict RNA-seq expression from pathological images and identify tumors with microsatellite instability [17]. Huang et al. developed the HistCode model to predict differential expression of tumor driver genes [18].

This paper proposed a deep learning framework based on weakly supervised contrastive learning, with the aim of establishing the links between molecular characteristics and histopathological morphology. We first split WSIs to millions of small tiles, and then performed contrastive learning pre-training on the large-scale unlabeled tiles to extract tile-level embeddings. For weakly supervised learning without annotation, the tile-level embeddings were aggregated using a gated attention mechanism to obtain slide-level feature, which is subsequently used in downstream whole-slide-level tasks. To validate the effectiveness of our method, we leveraged the computational pathological feature to infer gene expressions, molecular subtypes, treatment responses and clinical outcome. The extensive experiments demonstrated that our method can accurately identify tumor regions using only weakly supervised labels and exhibits exceptional performance in predicting breast cancer-related gene expression levels, as well as multiple clinically relevant tasks. Furthermore, spatial heatmaps generated based on the tile-level attention scores learned in tumor diagnosis and gene expression inference tasks exhibit high consistency with pathologist annotations and spatial transcriptomics. This indicates that the deep learning-based model effectively establishes the genotype-phenotype links, greatly enhancing the potential of digital pathology in clinical applications.

2 Results

2.1 Weakly supervised contrastive learning framework

Our study proposed a computational pathology framework that integrated self-supervised contrastive learning and attention-based multiple-instance aggregation to enable multi-task weakly supervised tasks based on the whole-slide images (WSIs). The framework comprised three stages (Fig. 1). In the first stage, super high-resolution whole slide images were split into small 256 x 256px tiles. The tiles without enough detectable foreground tissue by redefined area threshold were discarded. The second stage run contrastive learning-based pretraining on the large-scale unannotated tiles to obtain informative tile-level features. In the three stage, the gated attention mechanism was employed to aggregate tile-level features into slide-level representation. For each downstream tasks, the image encoder is fine-tuned and the attention score of each tile were learned based on weakly supervisory signals.

Unlike previous studies, we utilize self-supervised contrastive learning pretraining to extract features from millions of tiles. Contrastive learning seeks to minimize the distance between original and augmented images while simultaneously maximizing their distance from negative samples within a minibatch. This approach encourages the encoder to capture the essential feature of pathology images, resulting in more expressive embedding representations. Our previous work [18] and experiments conducted in this study (see Subsection 4) strongly supported that pre-training based on contrastive learning significantly enhances the predictive ability in downstream tasks.

Notably, our approach did not require pixel-level or region-of-interest (ROI) annotations. In fact, our model considered all tiles in the tissue regions of a WSI, and learned an attention score to each tile that convey its contribution or importance to the collective slide-level representation for a specific downstream task. This interpretation of the attention score is reflected in the slide-level aggregation rule of attention-based pooling, which computes the slide-level representation as the average of all patches in the slide weighted by their respective attention score. The tile-level attention score enable spatial localization of tumor subregions and gene expressions by spatial deconvolution of tiles to original slide (see Subsection 6 and7).

2.2 Accurate tumor diagnosis and tumor region localization using weakly supervisory signals

We first evaluated the proposed model on the TCGA-BRCA cohort by conducting a tumor diagnosis task. This classification task utilized only samples from fresh frozen tissue sections, comprising a total of 1,807 pathological slides, including 1,566 tumor tissues and 241 normal tissues. For objective performance assessment of our model, we randomly allocate 30% WSIs as an independent test set, and the remainder as training set, with the hyper-parameters fine-tuned by 3-fold cross-validation. Given that we only had

slide-level labels, we employed weakly supervised learning for tumor diagnosis and attempted to localize tumor regions. Specifically, each WSI was treated as a set composed of tiles. The tile features obtained in the pretraining stage were aggregated into slide features through attentive pooling and subsequently used to predict whether or not the slide contained tumor tissue. On the TCGA-BRCA cohort, our method achieved an accuracy of 0.99 and an ROC-AUC value of 0.99. Figure 2 (a-b) displays the ROC curve and confusion matrix on the test set. This result strongly supported that our model can accurately learn to distinguish whether a WSI contains tumor tissue using only weak supervisory signals. By performing UMAP dimensionality reduction on the latent representation of the slides, we discovered that tumor and non-tumor slides are significantly separated in the embedding space.

Since the slide features for tumor diagnosis are obtained by aggregating the features of tiles through an attention mechanism, tiles with high attention scores imply tumor regions. To assess whether our model can effectively locate tumor regions, we normalized the tile attention scores learned in the tumor diagnosis task and spatially deconvoluted each tile to generate heatmaps. Concurrently, we invited a pathologist with 10 years of experience to manually annotate the tumor regions. As depicted in Figure 2(d), the high-attention regions exhibit high consistency with the tumor areas annotated by the pathologist. Several selected tiles corresponding to high and low attention tiles display evident differences in tissue morphology.

2.3 Accurate prediction of gene expression relevant to breast cancer recurrence

To confirm that our model learned real and informative features from pathological images, we constructed an MLP model to predict the expression levels of 21 genes that have been confirmed to be closely related to breast cancer recurrence in previous studies [13]. Unlike existing studies that built one model per gene [19], we adopted multi-task learning to quantitatively predict the expression levels of 21 genes simultaneously. For this purpose, we downloaded RNA-seq data (FPKM-UQ) with matched WSI from TCGA-BRCA cohort, and obtained a total of 1,572 paired samples. We randomly selected 20% used as an independent test set and the remaining as training set, with hyperparameters tuned by 4-fold cross-validation. For this regression task, we used the correlation coefficient as the performance evaluation metric. Our method achieved an average correlation coefficient of 0.36, and the histogram of predicted correlation coefficients for 21 genes is shown in Figure 3(a). Further, we examined the top 10 well-predicted genes and found that most of them were related to cell proliferation. For example, MKI67 (r=0.504) is a well-known cell proliferation marker expressed in both tumor and non-tumor cells. Its overexpression has been found to be associated with tumor growth, histological staging, and tumor recurrence[20–22]. ESR1 (r=0.555) mutation is an important cause of endocrine resistance and acts as a clinical biomarker for metastatic hormone receptor-positive breast cancer[23]. AURKA (r=0.58)

belongs to the serine/threonine kinase family and its activation is essential for regulating cell division through mitosis and has been shown to be associated with prognosis in ER-positive, EBBR2-negative and TN breast cancer subtypes [24]. ERRB2 (r=0.252) overexpression/amplification is associated with malignant tumors and poor prognosis in breast cancer and is an important therapeutic target for breast cancer[25].

2.4 External cohort validated the generalizability of gene expression prediction

Unlike studies that built one prediction model per gene, our multi-task learning model demonstrates superior robustness. To further evaluate the generalizability of our model in predicting gene expression levels, we employed an independent external cohort from the CPTAC-BRCA project, consisting of 97 samples with matched pathological images and RNA-seq data. Notably, instead of training and testing on the external dataset, we completed model training on the TCGA-BRCA cohort and subsequently tested it on the CPTAC cohort to confirm that our model captured authentic features from a large number of tiles. Our model achieved an average correlation coefficient of 0.26, with the highest correlation coefficient for the MMP11 gene reaching 0.54 ($p<0.01$), as illustrated in Figure 3(b). These results indicate that deep learning-based gene expression prediction from pathological images can be generalized to external populations. The experimental results demonstrated that our method has made a significant progress towards practical clinical application.

2.5 cross-cohort validation demonstrate computational pathology highly indicative of recurrence risk

We further applied computational pathological features to predict the five-year recurrence of breast cancer. We evaluated the performance of our model on two datasets: the TCGA-BRCA cohort (n=272) and the CZCH cohort (n=91). A case was labeled as 1 if it had recurrence, metastasis, or a new tumor event within five years, and 0 otherwise. We applied three-fold cross-validation and AUC as the evaluation metrics. Figure 4(a) shows that our model achieved an AUC of 0.71 on the independent test set randomly sampled from the TCGA-BRCA cohort. To further test the generalizability of our model, we predicted patient recurrence events in the CZCH cohort using the model trained on the TCGA-BRCA cohort. We consider this cross-cohort prediction experiment as a strong validation of our model robustness. Figure 4(b) illustrates that the model attained an AUC of 0.69. These results indicate that our model has high effectiveness and generalizability in predicting breast cancer recurrence.

2.6 Accurate prediction of breast cancer molecular subtypes

PAM50 is a molecular subtyping method that classifies breast cancer into five subtypes (luminal A, luminal B, Her2-enriched, basal-like, and normal-like) based on the expression profiles of 50 breast cancer-related genes [11]. These subtypes have distinct molecular characteristics, prognoses, clinical behaviors, and treatment responses [26, 27]. For instance, luminal A and B subtypes are associated with relatively favorable outcomes and are mainly treated with surgery, chemotherapy, and endocrine therapy. Basal-like subtype is characterized by poor survival and lack of effective treatment options. Her2-enriched subtype requires targeted drugs or radiotherapy before surgical resection. In this study, we developed a model to predict PAM50 subtypes using pathological image features. Figure 5(a-b) shows that our model achieved high performance in predicting all five subtypes, with an average ROC-AUC value of 0.875 and PR-AUC value of 0.711. Notably, the AUC for the basal-like subtype reached 0.95. Our model can provide clinically actionable results by accurately predicting PAM50 subtypes using routine pathological images, especially for the basal-like subtype, which is otherwise costly and inaccessible in clinical practice.

2.7 Computational histopathological feature builds an independent prognostic factor

We developed a risk model to assess patient prognosis using a Cox proportional hazards deep neural network [28]. The model incorporated slide-level pathological image features, gene expressions, clinical data, PAM50 molecular subtypes, and other potential factors. We constructed the prognosis risk model using negative log-likelihood as the loss function and predicted overall survival and progression-free survival. The predictive performance was evaluated using the average concordance index (C-Index) on three-fold cross-validation. Kaplan-Meier curves were used to visualize the stratification quality between low-risk and high-risk patients. In the experiment with overall survival as the endpoint, our model achieved a concordance index of 0.685 on the test set. Figure 6(a) shows that the high-risk group had a significantly lower survival time than the low-risk group ($p<0.001$). A multivariate Cox regression analysis of the predicted risk score and other clinical data was conducted and displayed in a forest plot in Figure 6(b). The risk score based on pathological features was found to be a significant independent factor affecting overall survival (HR=3.205, $p<0.001$). In the experiment with progression-free survival as the endpoint, our model achieved a concordance index of 0.647 on the test set. Figures 6(c-d) illustrate that the high-risk group had a significantly higher prognosis risk than the low-risk group ($p<0.001$), and a multivariate Cox regression analysis indicated that the risk score constructed by pathological features was a significant independent factor affecting progression-free survival (HR=4.695, $p<0.001$).

2.8 Model ablation showed contrastive learning and attentive pooling improve performance

To validate that the features extracted through contrastive learning-based pre-training enhance the performance of downstream tasks, we conducted a comparative analysis of two contrastive learning algorithms - AdCo [29] and MoCo v2 [30], using ResNet50 [31] as the benchmark feature extractor. Our evaluation metrics included the Spearman correlation coefficient and Pearson correlation coefficient [29]. Figure 7(a) illustrated the performance of computational pathological features extracted by different algorithms in predicting gene expression levels. The experimental results indicated that the features extracted by AdCo contrastive learning yielded the most optimal performance, with an average Pearson and Spearman correlation coefficient of 0.41 and 0.37, respectively. Furthermore, we observed that both contrastive learning-based features significantly outperformed ResNet50 in this task.

We also examined the influence of various feature aggregation strategies on predictive performance by testing max pooling, average pooling [30], and attentive pooling on gene expression levels. As depicted in Figure 7(b), attentive pooling achieved performance far superior to the other two feature aggregation methods.

2.9 Spatial transcriptomics validate spatial deconvolution of gene expression

Spatial transcriptomics combines microscopic imaging and RNA sequencing technology to capture gene expression data while maintaining the spatial location of cells within the tissue. This allows the mapping of RNA-seq data onto tissue sections, enhancing the comprehension of the spatial organization of cells and their gene expression patterns [32]. To verify our model's capacity to perform spatial deconvolution, we produced heatmaps colored by tile-level attention scores of two biomarker genes, ERRB2 and MKI67. Meanwhile, we obtained spatial transcriptomics data from 10x Genomics and illustrated the spatial landscapes of these two gene expressions on the pathological image. As shown in Figure 8(a-b), we observed that the attention heatmaps of ERRB2 and MKI67 were highly consistent with their spatial expression landscapes derived from the spatial transcriptomics.

2.10 Effectively predicting trastuzumab treatment outcome

Our final analysis evaluated whether our model can capture pathological image features to predict the treatment response of breast cancer patients to trastuzumab. We obtained a cohort of patients treated with trastuzumab from Yale University. This cohort included the patients with a pre-treatment breast core biopsy with HER2 positive invasive breast carcinoma who then received neoadjuvant targeted therapy with trastuzumab. A case was designated as

responder if the pathological examination of surgical resection specimens did not report residual invasive, lympho-vascular invasion or metastatic carcinoma, and otherwise non-responders. After removal of the H&E images without enough detectable foreground tissue by redefined area threshold, we obtained 75 cases (34 responder and 41 non-responders).

Given that trastuzumab is a monoclonal antibody targeting HER2, we fine-tuned the feature encoder using HER2 level measured by Reverse-Phase Protein Array (RPPA) assay obtained from TCGA. This enabled our model to capture morphological features indicative of HER2 expression from pathological images. We employed five-fold cross-validation and used the ROC-AUC to evaluate model performance. Our results, depicted in Figure 9(a), demonstrated that our model achieved an AUC of 0.72 (95% CI: 0.65-0.81). In addition, we compared heatmaps generated by our model with tumor regions annotated by pathologists at Yale University and observed a high degree of consistency between the them, as shown in Figure 9(b). These findings support the feasibility of using pathological images to predict treatment response to trastuzumab, and demonstrate the effectiveness of deep learning models in identifying tissue morphological features associated with specific drug responses.

3 Materials and methods

3.1 Whole Slide Image

Whole slide images stained with HE provide a detailed view of cellular components, allowing for easy differentiation between cell nuclei and cytoplasm. We obtained breast cancer whole slide images from TCGA-BRCA cohort for model training. The TCGA-BRCA cohort contains 1,979 fresh frozen WSIs and 1,131 formalin fixed paraffin-embedded (FFPE) WSIs from 1,094 breast cancer patients. The slide-level diagnosis provided by TCGA database were used as ground truth for classification labels. The PAM50 molecular subtyping was also obtained from the TCGA database.

Three other independent cohort were included for model evaluation. The first was the CPTAC-BRCA cohort that includes 642 fresh frozen WSIs from 134 patients of breast cancer. It was used to validate our model performance to infer gene expression levels using extracted pathological image features. The second independent cohort came from the Changzhou Cancer Hospital (CZCH) in Jiangsu, China. The CZCH cohort collected 91 FFPE WSIs from 91 breast cancer patients, including 15 recurrence cases within 5 years. The CZCH cohort was used as an external validation set for recurrence prediction.

The third is the Yale trastuzumab response cohort that was used to evaluate model ability in predicting drug response. The Yale cohort contained 75 FFPE WSIs of 75 cases of breast cancer (34 responder and 41 non-responders). A case was designated as responder if the pathological examination of surgical resection specimens did not report residual invasive, lympho-vascular invasion

or metastatic carcinoma, and otherwise non-responders. Note that only the slides with a magnification greater than 20* were included in this study.

3.2 Slide annotation

A pathologist with more than ten-year experience was asked to manually annotate a few slides. The annotated slides were used for the validation of spatial localization of tumor and necrosis regions. During annotation, the pathologist was blinded with regard to any molecular or clinical feature.

3.3 RNA-seq dataset

We collected the RNA-seq data matched to whole slide images from TCGA-BRCA and CPTAC-BRCA cohorts to train and validate our model capability in predicting gene expression levels, respectively. We adopted the upper quartile normalized (FPKM-UQ) fragments per kilobase per million mapped fragments (FPKMs) as expected values to train the model. Since the range of gene expression levels spans several orders of magnitude, straightforward regression on the raw RNA-Seq data would cause the model to focus only on the most strongly expressed genes, neglecting the signals of other genes. To overcome this problem, we applied a log10(1+x) transformation to the gene FPKM-UQ expression values [17].

3.4 Spatial transcriptomic data

We acquired spatial transcriptomic data from a breast cancer specimen using the Visium Spatial Gene Expression protocol from 10x Genomics. To explore the spatial locations of gene expression, we utilized the 10x Loupe Browser, a desktop application that offers interactive visualization capabilities for various 10x Genomics solutions. This enabled us to visualize the spatial landscape of specific biomarker genes experssion across the pathology image.

3.5 Preprocessing of whole-slide images

Due to the ultra-high resolution of pathological images, it is inhibitive to directly use them as inputs of deep neural networks. Therefore, we tessellated WSIs into small-size squares (tiles) so that they were ready as input of deep learning model. For each slide, we use OpenSlide to read it into memory and then convert RGB to the HSV color space. For tissue regions (foreground), we generate a binary mask by thresholding the saturation channel in the HSV color space, smoothing edges and performing morphological closure to fill small gaps and holes. This effectively segments the WSI image into tissue and non-tissue regions. After segmentation, each WSI was split into 256×256 pixel tiles within the foreground region and filter out white image tiles without enough tissue (pixel background value >200). It is worth noting that in order to improve access efficiency, we save only the coordinates and metadata of each tile in hdf5 format.

3.6 Contrastive learning for feature extraction

Due to the scarcity of pixel-level labels of pathological images, traditional fully supervised learning is not applicable in our study. We used self-supervised contrastive learning on large-scale unlabeled tiles to learn latent representations. In our study, we adopted the Adversarial Contrastive learning (AdCo) algorithm for pretraining. AdCo stands for adversarial contrastive for efficient learning of unsupervised representations from self-trained negative adversaries. It relies on constructing a collection of negative examples that are sufficiently hard to discriminate against positive queries when their representations are self-trained. We also test another contrastive learning algorithm , MoCo v2, which is an improved version of the Momentum Contrast self-supervised learning algorithm. A ResNet50 network was used to benchmark the performance of contrastive learning.

For contrastive learning, the positive pairs were generated through data augmentation techniques such as image rotation, random cropping, and color transformation. Formally, denote the input tile as by x_i, we used the ResNet50 CNN network as the backbone encoder, assuming that the encoder f transforms the tile into a latent representation $h_i = f(x_i)$, which is then projected onto an embedding q_i through multiple fully-conneted layers. A negative sample queue is initialized to store the latent representations of the tiles. Contrastive learning trains the network parameters θ to distinguish between the query sample q_i and its positive partner from a set of negative samples K. The contrastive loss is defined as:

$$L = \frac{-1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{q}_i^\top \mathbf{q}_i'/\tau\right)}{\exp\left(\mathbf{q}_i^\top \mathbf{q}_i'/\tau\right) + \sum_{k=1}^{K} \exp\left(\mathbf{q}_i^\top \mathbf{n}_k/\tau\right)} \tag{1}$$

where q_i and $q_i^{'}$ represent the embeddings of a parir of positive samples, n_k refers to the embedding of a negative sample and τ is the temperature parameter.

Inspired by adversarial learning, AdCo aims to generate challenging negative samples to be distinguished from query samples. It extended the memory bank that was used to stack negative sample embeddings by regarding negative samples as learnable weights. This allows the learned negative samples to closely track changes in the learned embedding representations. AdCo takes the embedding of all samples as negative adversaries to network θ and updates the negative samples to maximize the contrastive loss, so that the adversarial negative samples were push closer to query samples. So, we have the following min-max problem:

$$\theta^*, \mathcal{M}^* = \arg\min_{\theta} \max_{\mathcal{M}} L(\theta, \mathcal{M}) \tag{2}$$

in which M represents the set of dynamically updated adversarial negative samples. In fact, M can also be considered as a set of model parameters. During the training process, the network parameter θ is updated along the direction of gradient descent, while M is updated along the direction of gradient ascent.

Thus, the parameters θ and M are alternately updated as follows:

$$\theta \leftarrow \theta - \lambda_\theta \frac{\partial L(\theta, \mathcal{M})}{\partial \theta} \tag{3}$$

$$m_k \leftarrow m_k - \lambda_{m_k} \frac{\partial L(\theta, \mathcal{M})}{\partial m_k} \tag{4}$$

Once the contrastive learning-based pre-training is completed, we obtained the low-dimensional embedding of each tile using the trained ResNet50 backbone encoder. As a result, we converted each 256*256px tile into a 1024-dimensional vector. The extracted features were stored using hdf5 ready for downstream tasks. Note that we actually froze the encoder parameters during the downstream tasks.

3.7 Attentive pooling for tile-level feature aggregation

Since the downstream task is based on slide-level features, we aggregate tile-level features into slide-level features. Unlike max pooling or average pooling, we use gated attention pooling here. Suppose a slide is divided into n tiles S=$\{p_1, p_2, ...p_n\}$, the gated attention pooling is essentially a weighted average pooling where the weight is trainable.

$$S = \sum_{i=1}^{N} a_i h_i \tag{5}$$

in which a_i is the learnable parameter indicating the importance of i-th tile to specific task. Formally, a_i is defined as below:

$$a_i = \frac{exp(w^T(tanh(Vh_i) \bullet sigm(Uh_i)))}{\sum_{j=1}^{N} exp(w^T(tanh(Vh_j) \bullet sigm(Uh_j)))} \tag{6}$$

where U and V are all learnable parameters, $\bullet$ is an element-wise multiplication and sigm() is the sigmoid activation function. and wT is used to transfer the feature vector into raw attention score. The major difference between gated-attention pooling and max/mean pooling was that it introduced learnable parameters that were iteratively updated during the training process, making the model highly flexible and adaptable to different downstream tasks. More importantly, the attention weights after training reflected the importance of tile-level features to downstream task, which makes the model interpretable.

3.8 Model training and inference

During the contrastive learning phase, we randomly select a fraction of tiles from each slide and stack them in the memory bank, since the large number of tiles exceeds the memory capacity. This is a reasonable approach because individual tiles have relatively low information content compared to natural

images, so that a certain proportion of randomly extracted tiles enables us to obtain the overall information of the slide. The adversarial negative samples were dynamically updated to force the encoder to capture the essential features of each tile and distinguish it from others.

In our implementation, ResNet50 was used as the backbone network. We only used its first four layers (the output of the last layer is 1024) and load the weights pre-trained on ImageNet. Two fully connected layers are appended to the backbone network to project the. SGD[33] is used as the optimizer. The learning rate of the backbone network is set to 0.03. Weight decay is set to 0.0001 and momentum is set to 0.9.

3.9 Downstream tasks

We applied transfer learning to leverage slide-level features for various downstream tasks, such as tumor diagnosis, gene expression level estimation, PAM50 molecular subtyping, recurrence and drug response prediction.

We performed tumor diagnosis on the TCGA-BRCA cohort, using the slide label (tumor or non-tumor) provided by the TCGA database as the ground truth. We employed a fully-connected layer plus a softmax layer as the prediction model that took the slide feature as input. The cross-entropy was used as the loss function for the classification task. Denote by y and y' the true and predicted label for tumor diagnosis, the loss function is as below:

$$\mathrm{L} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i} \log \mathrm{y'_i} + (1 - \mathrm{y_i}) \log (1 - \mathrm{y'_i}) \tag{7}$$

The gene expression level estimation was formulated as a regression task. Taking as input the extracted slide features, we used a multi-task learning model with a fully-connected layer and a output layer to predict the gene expression level. For instance, the output layer had 21 nodes corresponding to the expression levels of 21 genes associated with breast cancer recurrence. In our practice, we have test multiple hidden layers, and found that a single hidden layer could achieve superior performance in the regression task. We used the mean squared error as the loss function:

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i - y'_i)^2 \tag{8}$$

For PAM50 molecular subtyping, it was formulated as a mutli-class classification problem. The model consisted of a fully-connected layer and a softmax layer. The four output nodes correspond to four molecular subtypes (luminal A, luminal B, Her2-enriched, and basal-like).

For recurrence, a case was labeled as 1 if it had recurrence, metastasis, or a new tumor event within five years, and 0 otherwise. For drug response, a case

was designated as responder if the pathological examination of surgical resection specimens did not report residual invasive, lympho-vascular invasion or metastatic carcinoma, and otherwise non-responders. Therefore, the prediction of recurrence and drug response from pathological features were both formulated as binary classification problems, and the cross-entropy loss function was used.

During the training process, we used the Adam optimizer with a weight decay of 0.0001 and an initial learning rate of 0.003. We applied a decay strategy to dynamically adjust the learning rate. The learning rate was multiplied by 0.1 every time the model trained for 10 epochs.

3.10 Establishment of prognosis risk model

To build a prognosis risk model based on slide-level features, we employed a deep learning-based Cox proportional hazard regression model. The formula of the risk function is as follows:

$$h(t|x) = \lim_{\triangle t \to 0} \frac{Pr(t \leq O \leq t + \triangle t | O \geq t;\ x)}{\triangle t} \tag{9}$$

It estimates the instantaneous death rate of individual x at time t. The Cox proportional hazard regression models it as:

$$h(t|x) = h_0\,(t)\, e^{\hat{h_\theta}(x)} \tag{10}$$

where x is the input feature that represents the baseline risk function at time t. In this study, we concatenate the slide features, clinical data, and the 21 gene expression levels into a one-dimensional vector as the model input. We built a multilayer perceptron model composed of five fully-connected layers. The network output was a single node using the selu activation function, which estimated the log hazard function in the COX model. We trained the network by setting the objective function to the regularized average negative log partial likelihood as below:

$$L_\theta = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h_\theta}(x_i) - log \sum_{j \in R(t_i)} e^{\hat{h_\theta}(x_j)})) + \lambda \|\theta\|_2^2 \tag{11}$$

In which $R(t_i)$ is the risk set at time t_i indicating the set of patients who are at risk at time t, $E_i = 1$ indicates that the observed event (death) occurred, $N_{E=1}$ is the number of samples with observed events, and λ is the ℓ_2 regularization parameter. During the training process, Adam was used as the optimizer, the learning rate was set to 0.01, the weight decay was set to 0.001 and the regularization parameter was set to 0.15.

4 Discussion and Conlusion

In this paper we propose a neural network model based on contrast learning and attention mechanism for tumor diagnosis, gene expression prediction and recurrence risk prediction on whole pathology images. Our experimental results show that our model achieves good performance on tumor diagnosis task and recurrence prediction task, and is able to significantly predict gene expression levels from WSIs. In addition, we generate spatially visualized heat maps from attention scores generated by the attention mechanism and compare them with pathologist annotations and spatial transcriptomes. We conclude that introducing contrast learning into the feature extraction stage greatly improves the problem of no effective annotation of data in the pathology image domain, and significantly improves a range of downstream tasks. In the future, we hope to use more contrast learning frameworks to fuse different multi-scale features to effectively improve our model and test our approach on more tumor datasets.

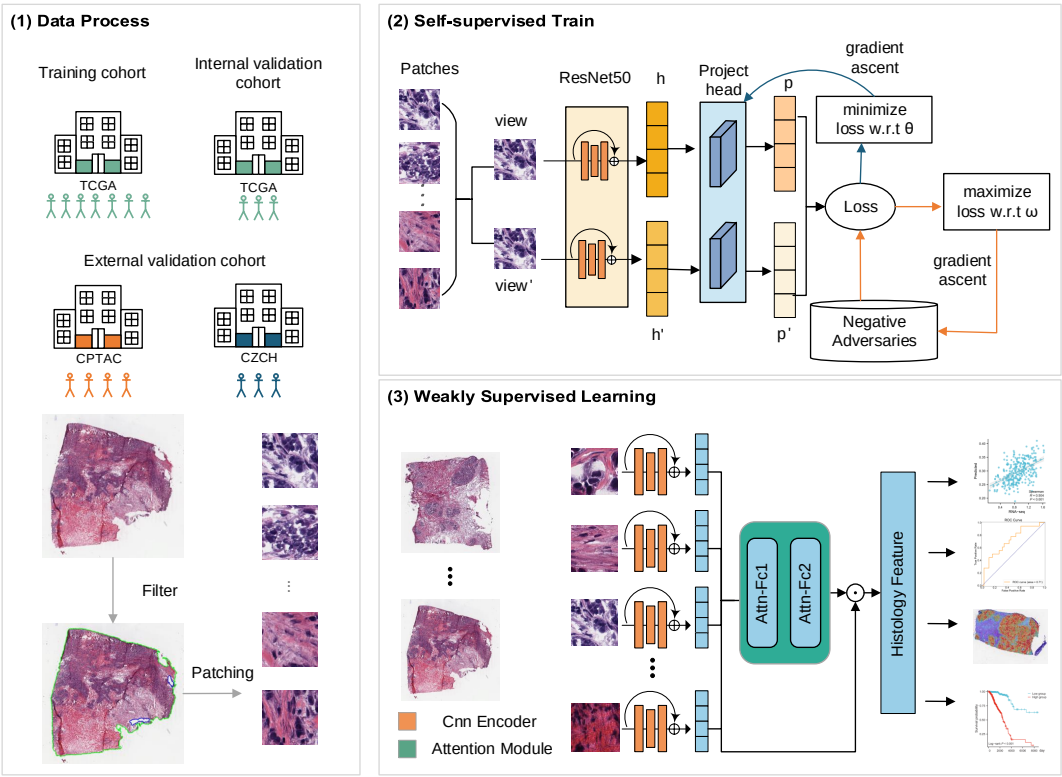
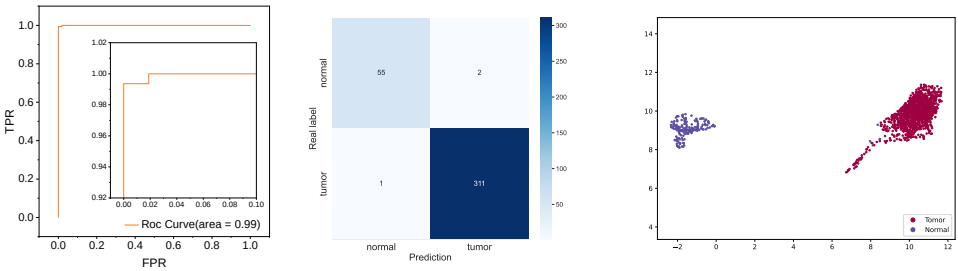
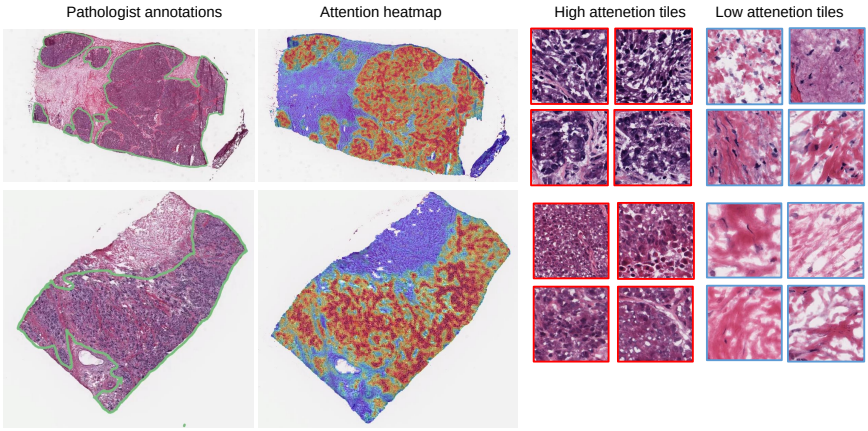


Fig. 1: Illustrative flowchart of the framework. There were three steps, slide proprocessing, contrastive learning-based pretraining and downstream tasks.



(a) (b) (c)



(d)

Fig. 2: Figure a shows the ROC curve of the predicted results of our model on the test set;Figure b shows the Confusion matrix of the predicted results of our model on the test set;Figure c shows the slide-level feature umap downscaled graph learned by our model;Figure d demonstrates the spatial localization of the tumor region as implied by the attention mechanism compared to the pathologist's annotation. The heatmap (middle) was generated by spatially deconvoluting the original slides, with each slide colored according to its normalized attention score. The areas circled in green are areas of tumor and necrosis annotated by experienced pathologists (left). The right column shows some representative tiles with higher and lower attention scores.

Declarations

- Funding This work was supported by National Natural Science Foundation of China (62072058, 82073339).

- Conflict of interest
  The authors declare no competing interests.

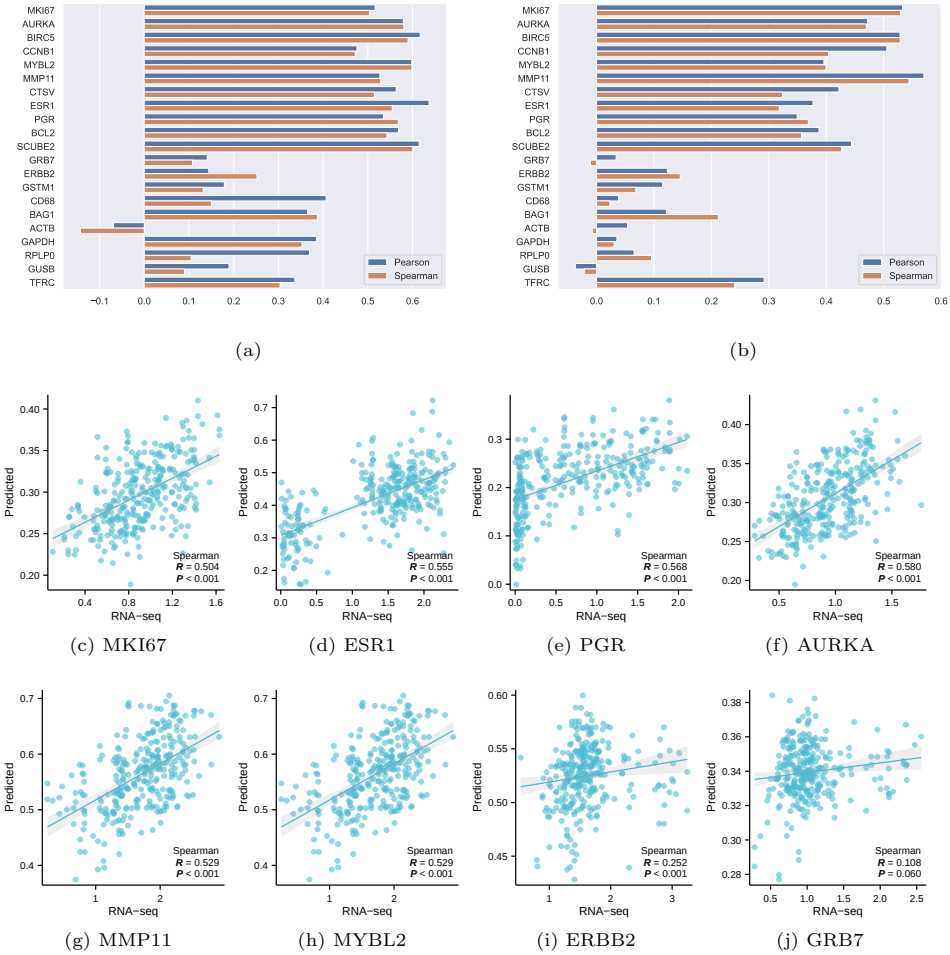
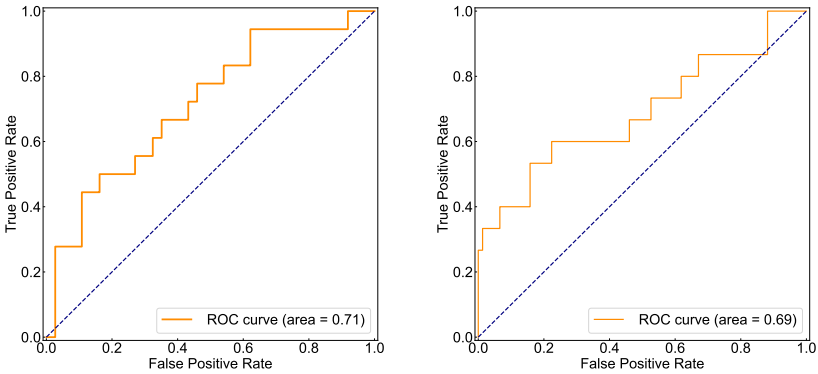


Fig. 3: Figure a shows the 21-gene predicted correlation coefficient histogram on the tcga test set;Figure a shows the 21-gene predicted correlation coefficient histogram on the cptac independent external validation set.Figure c-j gives a scatter plot of some of the predicted gene expressions associated with breast cancer as predicted by the model.

- Ethics approval
- Data availability
  The TCGA-BRCA diagnostic and frozen slides with corresponding labels, and the matched RNA-seq data are available from the NIH genomic data commons (https://portal.gdc.cancer. gov). The CPTAC-BRCA whole-slide with corresponding labels, and the matched RNA-seq data are available from the NIH cancer imaging archive (https://cancerimagingarchive.net/



(a) (b)

Fig. 4: Figure A shows the ROC curve of the predicted results of our model on the test set; Figure B shows the ROC curve of the predicted results of our model on the external validation sets

datascope/cptac). The whole slide images and drug response data from Yale trastuzumab response cohort are available at TCIA database https://wiki.cancerimagingarchive.net/. The spatial transcriptomic data of a breast cancer specimen from 10X genomics are available at https://www.10xgenomics.com/. The All reasonable requests for academic use of in-house raw and analysed data can be addressed to the corresponding author.

- Code availability
- Authors' contributions
  H.L. and Y.Z. conceived the study and designed the experiments. Y.Z. performed the experimental analysis. J.L. curated the in-house datasets and collected smartphone microscopy data. H.L. and Y.Z. prepared the manuscript. H.L. and J.L. supervised the research.

References

[1] Wilkinson, L., Gathani, T.: Understanding breast cancer as a global health concern. British Journal of Radiology **95** (2022)

[2] Noorbakhsh, J., Farahmand, S., et al.: Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nature Communications **11** (2020)

[3] Wang, S., Rong, R., *et al.*: Computational staining of pathology images to study the tumor microenvironment in lung cancer. Cancer Research **80**, 2056–2066 (2020)

[4] Wang, Y., Acs, B., *et al.*: Improved breast cancer histological grading using deep learning. Annals of Oncology **33**, 89–98 (2022)

[5] Kather, J.N., Pearson, A.T., *et al.*: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine **25**, 1054–1056 (2019)

[6] Jain, M.S., Massoud, T.F.: Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nature Machine Intelligence **2**, 356–362 (2020)

[7] Saltz, J., Gupta, R., *et al.*: Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Reports **23**, 181–1937 (2018)

[8] Lu, M.Y., Williamson, D.F.K., *et al.*: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**, 555–570 (2021)

[9] Zhang, H., Meng, Y., et al.: DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification

[10] Wu, H., Wang, Z., et al.: Cross-patch Dense Contrastive Learning for Semi-supervised Segmentation of Cellular Nuclei in Histopathologic Images

[11] Bernard, P.S., et al.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology **27** (2009)

[12] Sun, D., Chen, J., et al.: Establishment of a 12-gene expression signature to predict colon cancer prognosis. PeerJ **2018** (2018)

[13] Paik, S., Shak, S., *et al.*: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine **351**, 2817–2826 (2004)

[14] Dowsett, M., Cuzick, J., *et al.*: Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: A transatac study. Journal of Clinical Oncology **28**, 1829–1834 (2010)

[15] Sparano, J.A., Gray, R.J., *et al.*: Prospective validation of a 21-gene expression assay in breast cancer. New England Journal of Medicine **373**, 2005–2014 (2015)

[16] Kather, J.N., Heij, L.R., *et al.*: Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer **1**, 789–799 (2020)

[17] Schmauch, B., Romagnoni, A., et al.: A deep learning model to predict rna-seq expression of tumours from whole slide images. Nature Communications **11** (2020)

[18] Huang, H., Zhou, G., et al.: Pathology Image-based Differential Gene Expression Contrastive Learning-based Computational Histopathology Predict Differential Expression of Cancer Driver Genes

[19] Wang, Y., Kartasalo, K., *et al.*: Predicting molecular phenotypes from histopathology images: A transcriptome-wide expression–morphology analysis in breast cancer. Cancer Research **81**, 5115–5126 (2021)

[20] Ng, I.O.L., Na, J., et al.: Ki-67 antigen expression in hepatocellular carcinoma using monoclonal antibody mib1. a comparison with proliferating cell nuclear antigen. American Journal of Clinical Pathology **104** (1995)

[21] SHIRABE, K., KITAMURA, M., et al.: A long-term survivor of ruptured hepatocellular carcinoma after hepatic resection. Journal of Gastroenterology and Hepatology **10** (1995)

[22] Luo, Y., Ren, F., et al.: Clinicopathological and Prognostic Significance of High Ki-67 Labeling Index in Hepatocellular Carcinoma Patients: A Meta-analysis

[23] Brett, J.O., Spring, L.M., et al.: ESR1 Mutation as an Emerging Clinical Biomarker in Metastatic Hormone Receptor-positive Breast Cancer

[24] Du, R., Huang, C., et al.: Targeting AURKA in Cancer: Molecular Mechanisms and Opportunities for Cancer Therapy

[25] Yarden, Y.: Biology of her2 and its importance in breast cancer, vol. 61 (2001)

[26] Caan, B.J., Sweeney, C., et al.: Intrinsic subtypes from the pam50 gene expression assay in a population-based breast cancer survivor cohort: Prognostication of short- and long-term outcomes. Cancer Epidemiology Biomarkers and Prevention **23** (2014)

[27] Coates, A.S., Winer, E.P., et al.: Tailoring therapies-improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. Annals of Oncology **26** (2015)

[28] Katzman, J.L., Shaham, U., et al.: Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology **18** (2018)

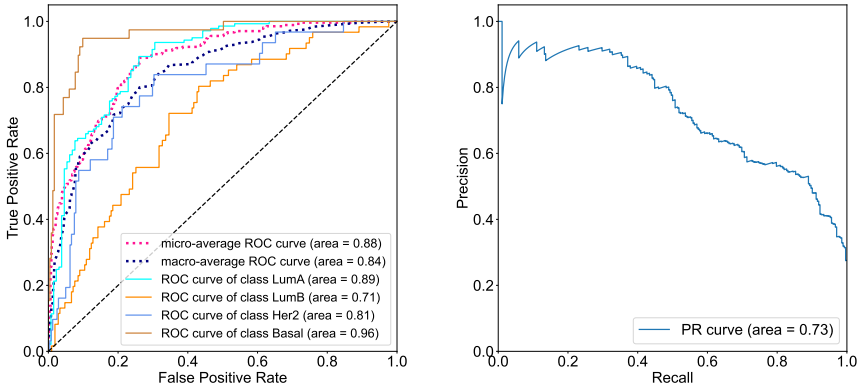
[29] Hu, Q., Wang, X., et al.: Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries (2020)

[30] Chen, X., Fan, H., et al.: Improved baselines with momentum contrastive learning (2020)

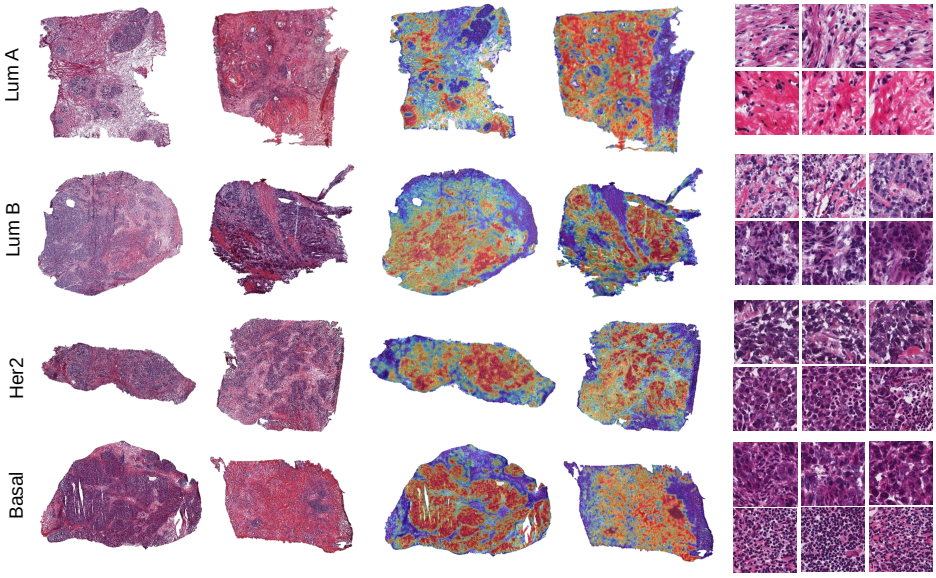
[31] He, K., Zhang, X., *et al.*: Deep residual learning for image recognition, vol. 2016-December (2016)

[32] Ståhl, P.L., Salmén, F., et al.: Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics

[33] Bottou, L.: Large-scale machine learning with stochastic gradient descent. (2010)



(a) (b)



(c)

Fig. 5: Figure A shows the ROC curve of the predicted results of our model on the test set; Figure B shows the PR curve of the predicted results of our model on the test set; Figure C shows the attentional heatmap and high attentional tiles of our method on the four molecular typologies.

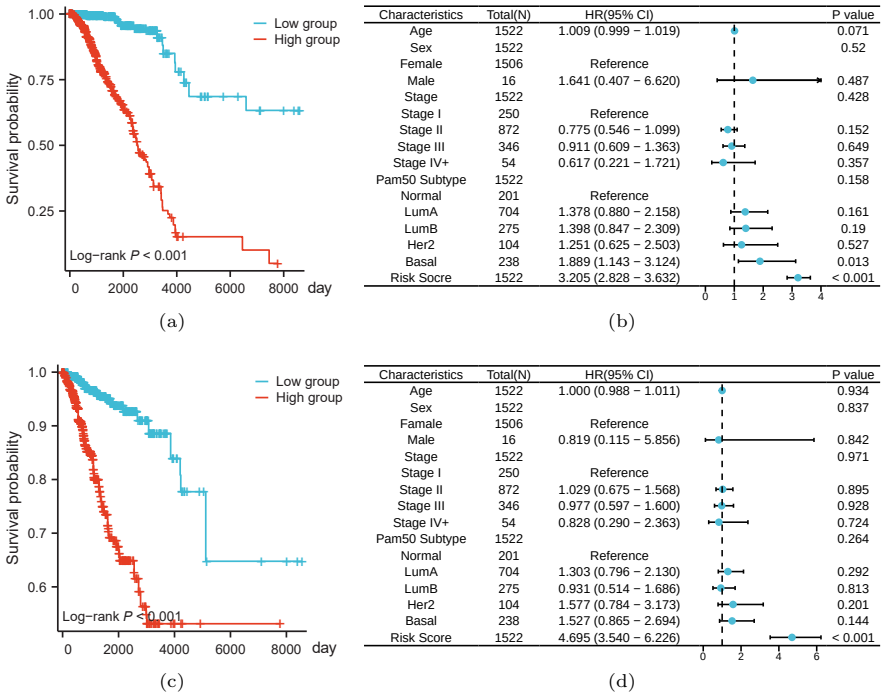
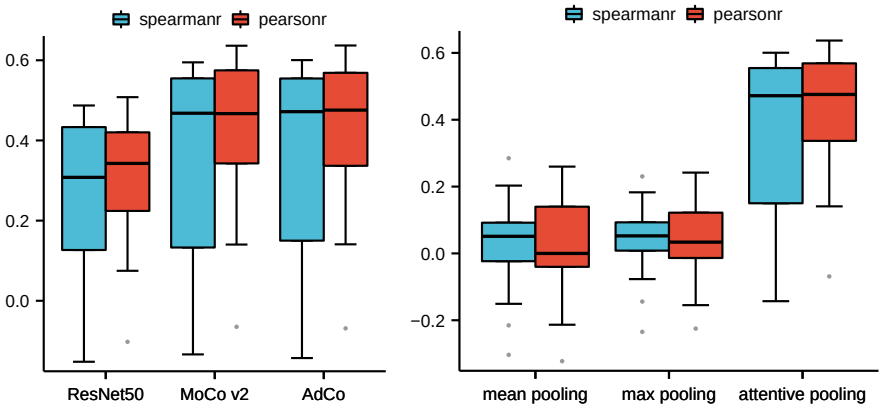
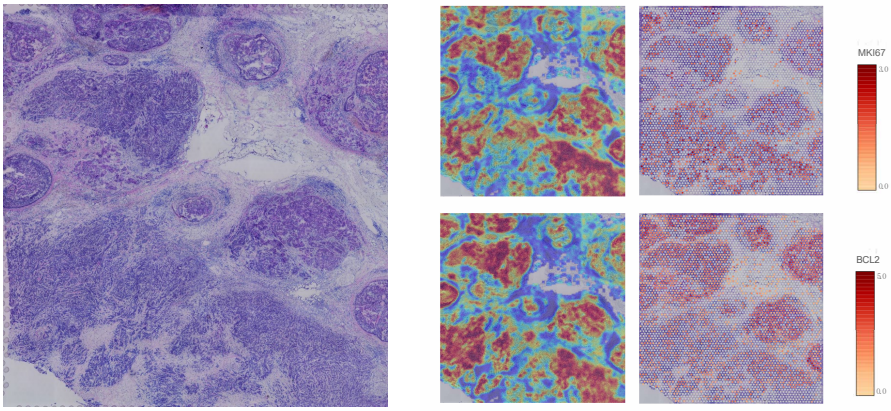


Fig. 6: Figure a shows the survival curves (P-value<0.05) constructed from the risk scores for the endpoint event of overall survival;Figure b shows the forest plot of OS univariate analysis on the TCGA-BRCA dataset for risk score, age, pathological stage and molecular subtype;Figure c shows the survival curves (P-value<0.05) constructed from the risk scores of the PFI for the endpoint events;Figure d shows the forest plot of PFI univariate analysis on the TCGA-BRCA dataset for risk score, age, pathological stage and molecular subtype.



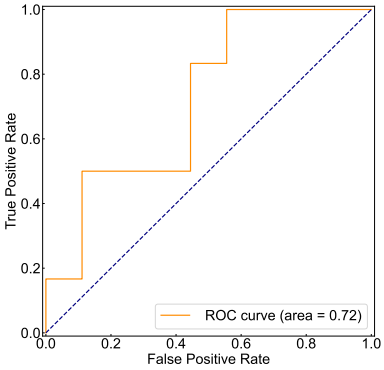
(a) (b)

Fig. 7: Figure A shows the prediction results of ResNet50, MoCo v2 and Adco respectively; Figure B shows the prediction results of mean pooling,max pooling and attention pooling.

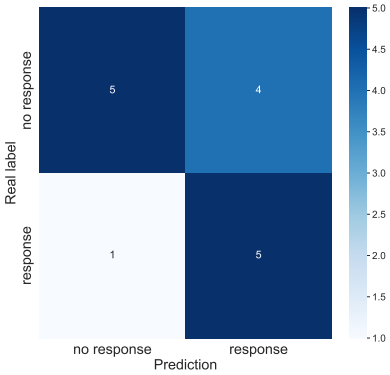


(a) (b)

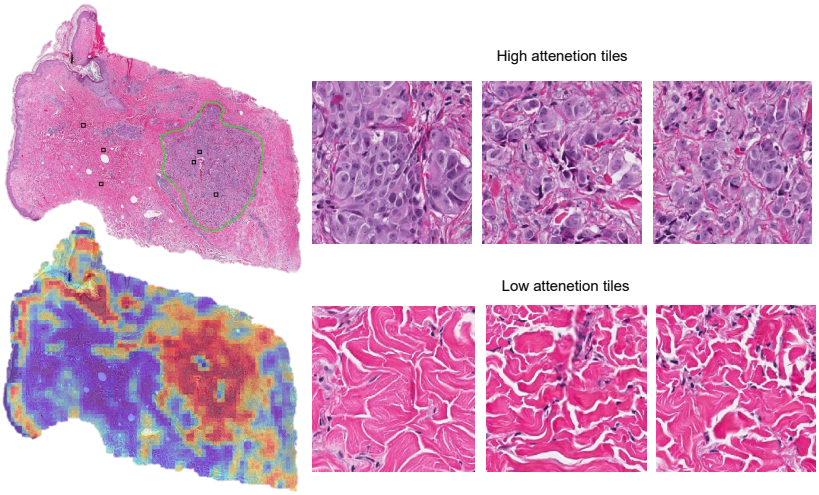
Fig. 8: Figure A shows the original pathological section; Figure B shows the comparison between the heat map generated by our model for MKI64 and BCL2 genes and the gene high expression regions localized by spatial transcriptome techniques.



(a)



(b)



(c)

Fig. 9: Figure A shows the ROC curve of the predicted results on the Yale trastuzumab response cohort; Figure B shows the Confusion matrix of trastuzumab response predictor;Figure C shows the attention heatmap and regions of high and low attention of trastuzumab response predictor.