BIG DATA MINING AND ANALYTICS
ISSN 2096-0654
DOI: 10.26599/BDMA.2024.9020012

A Disentangled Representation-Based Multimodal Fusion Framework Integrating Pathomics and Radiomics for KRAS Mutation Detection in Colorectal Cancer

Zhilong Lv, Rui Yan, Yuexiao Lin, Lin Gao*, Fa Zhang*, and Ying Wang*

Abstract: Kirsten rat sarcoma viral oncogene homolog (namely KRAS) is a key biomarker for prognostic analysis and targeted therapy of colorectal cancer. Recently, the advancement of machine learning, especially deep learning, has greatly promoted the development of KRAS mutation detection from tumor phenotype data, such as pathology slides or radiology images. However, there are still two major problems in existing studies: inadequate single-modal feature learning and lack of multimodal phenotypic feature fusion. In this paper, we propose a Disentangled Representation-based Multimodal Fusion framework integrating Pathomics and Radiomics (DRMF-PaRa) for KRAS mutation detection. Specifically, the DRMF-PaRa model consists of three parts: (1) the pathomics learning module, which introduces a tissue-guided Transformer model to extract more comprehensive and targeted pathological features; (2) the radiomics learning module, which captures the generic hand-crafted radiomics features and the task-specific deep radiomics features; (3) the disentangled representation-based multimodal fusion module, which learns factorized subspaces for each modality and provides a holistic view of the two heterogeneous phenotypic features. The proposed model is developed and evaluated on a multi modality dataset of 111 colorectal cancer patients with whole slide images and contrast-enhanced CT. The experimental results demonstrate the superiority of the proposed DRMF-PaRa model with an accuracy of 0.876 and an AUC of 0.865 for KRAS mutation detection.

Key words: KRAS mutation detection; multimodal feature fusion; pathomics; radiomics

1 Introduction

Colorectal cancer (CRC) is a common malignant disease that starts out as a precancerous polyp formed by abnormal growths of epithelial cells in the colon or rectum[1]. CRC is the third leading cause of cancer incidence and the second leading cause of cancer mortality worldwide, with approximately 1.9 million new cases and 935 000 deaths in 2020[2]. The International Agency for Research on Cancer (IARC) estimates that the burden of CRC will continue to

• Zhilong Lv and Lin Gao are with School of Computer Science and Technology, Xidian University, Xi'an 710071, China. E-mail: lvzhilong@xidian.edu.cn; lgao@mail.xidian.edu.cn.
• Rui Yan is with School of Biomedical Engineering, University of Science and Technology of China, Hefei 230026, China. E-mail: yanrui@ustc.edu.cn.
• Yuexiao Lin is with Department of General Surgery, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China. E-mail: linyuexiao2506@163.com.
• Fa Zhang is with School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail: zhangfa@bit.edu.cn.
• Ying Wang is with Department of Pathology, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China. E-mail: wangying_blk@126.com.
* To whom correspondence should be addressed.
Manuscript received: 2024-01-08; revised: 2024-01-27; accepted: 2024-02-28

increase, reaching 3.2 million new cases and 1.6 million deaths worldwide by 2040[3].

In the past decades, traditional treatments, including surgical resection, radiotherapy, and chemotherapy, have been the general choice for patients with CRC. However, because CRC is a highly heterogeneous tumor disease involving multiple carcinogenic pathways[4, 5], these one-size-fits-all treatments have failed to achieve consistent beneficial effects. Recently, precision medicine that aims to provide molecule-guided personalized therapy, has been gradually applied to the CRC patients in order to improve the effectiveness of treatment. The identification of molecular biomarkers that correlate with response to therapy or function in disease initiation and progression is fundamental to cancer precision medicine[6].

Kirsten rat sarcoma viral oncogene homolog (namely KRAS), a member of the RAS gene family, is one of the most frequently mutated oncogenes in CRC, accounting for more than 40% of CRC cases[7]. KRAS mutations are single nucleotide point mutations that occur predominantly at glycine in codons 12 and 13 of exon 2[8]. Clinical studies have demonstrated that KRAS-mutant CRC patients have a worse prognosis than wild-type patients, especially when CRC metastasizes to the liver or lung[9]. In addition, KRAS-mutant CRC patients show resistance to monoclonal antibodies against Epidermal Growth Factor Receptor (EGFR), such as cetuximab and panitumumab, which have been approved by the United States Food and Drug Administration (FDA) for the first-line treatment of CRC[10]. Therefore, KRAS has been considered as a critical prognostic and predictive response biomarker for CRC[11]. However, due to the long turnaround time and high cost of the molecular diagnostic techniques, such as Sanger sequencing, pyrosequencing, and allele-specific PCR, KRAS mutation detection has not yet been widely adopted in clinical practice[12]. There is an urgent need to develop a more convenient and efficient method for KRAS mutation detection.

Recently, the advancement of machine learning, especially deep learning, has greatly promoted the development of genotype-phenotype correlation research, providing a promising prospect for molecular marker detection from tumor phenotype data, including pathology slides and radiological images. In 2018, Coudray et al.[13] presented a deep learning based framework to effectively detect the six most common mutated genes, including KRAS, from non-small cell lung cancer pathology slides for the first time. At the same time, Yang et al.[14] used the ReliefF algorithm[15] to select three key features from 346 candidate hand-crafted radiomics features as input variables for the SVM model, which achieves KRAS mutation detection from CT images in colorectal cancer. The above studies have demonstrated the feasibility of KRAS mutation detection from pathology slides or radiology images, which sparks a boom in subsequent research. However, there are still two major problems with existing KRAS mutation detection studies. However, there are still two major problems with existing studies of KRAS mutation detection. (1) Inadequate single-modal phenotypic feature learning. The pathomics learning models based on random patch selection cannot capture the effective pathological features, and the radiomics learning models based on traditional feature engineering cannot sufficiently characterize the radiological features. (2) Lack of multimodal phenotypic feature fusion. The current studies generally focus on single-modality data and lack comprehensive integration of multimodal phenotypic feature information.

In this paper, we propose a Disentangled Representation-based Multimodal Fusion framework integrating Pathomics and Radiomics (DRMF-PaRa) for KRAS mutation detection. The main contributions of this work can be summarized as follows:

(1) we introduce a novel tissue-guided Transformer model for the pathomics learning to extract more comprehensive and targeted pathological features;

(2) we adopt the combination of the hand-crafted radiomics and the supervised contrastive learning-based radiomics in the radiomics learning to capture both the generic hand-crafted features and the task-specific deep features;

(3) we introduce the disentangled representation learning in the multimodal phenotypic feature fusion to learn factorized subspaces for each modality and provide a holistic view for KRAS mutation detection from heterogeneous multimodal data.

2 Related Work

2.1 KRAS mutation detection from phenotype data

In 2018, Coudray et al.[13] first presented a deep learning based framework to effectively predict the six most common mutated genes in non-small cell lung

cancer, STK11, EGFR, FAT1, SETBP1, KRAS and TP53, from Hematoxylin and Eosin (H&E)-stained slides. Inspired by this work, Jang et al.[16] applied the deep learning models to predict the five common and clinically relevant mutated genes in CRC, including KRAS, from the pathology patches with high tumor probability detected by a tumor tissue classifier. Jiang et al.[17] used the endoscopic knowledge to build a deep learning framework consisting of multiple models with the same backbone network to predict CRC subtypes and KRAS mutations. Then, Schrammen et al.[18] proposed a Slide-Level Assessment Model (SLAM) for simultaneous tumor detection and prediction of genetic alterations. It uses a single off-the-shelf neural network to predict molecular alterations directly from routine pathology slides without manual annotation, improving upon previous methods by automatically excluding normal and non-informative tissue regions. Ding et al.[19] proposed a graph neural network approach to emphasize the spatialization of tumor tiles for a comprehensive evaluation of predicting cross-level molecular profiles of genetic mutations, copy number alterations, and functional protein expressions from whole slide images. In addition, Wagner et al.[20] developed a Transformer-based pipeline for end-to-end biomarker prediction from pathology slides by combining a pre-trained Transformer network for patch aggregation, which achieved an AUC of 0.80 in KRAS mutation detection. However, due to the large image size and extremely high resolution of whole slide images, these existing studies generally rely on the random patch selection strategy and simple patch-level aggregation models for slide-level prediction, resulting in inadequate pathological representation and unstable performance.

In addition to pathology slides, radiology data represented by CT has also been shown to be useful for KRAS mutation detection[21]. As a preliminary study, Yang et al.[14] selected three key features from 346 candidate hand-crafted radiomics features as input variables for the SVM model to achieve an AUC of 0.829 in KRAS mutation detection. Then, Taguchi et al.[22] adopted a multivariate machine learning method with 14 comprehensive CT texture parameters to achieve a superior performance in predicting of the KRAS mutation status in CRC. Shi et al.[23] used a deep artificial neural network based on radiomics and semantic features to predict the mutation status of RAS and BRAF with an AUC of 0.79 in the validation cohort. Recently, deep learning, which can avoid complex feature engineering and automatically learn task-related deep features, has also been gradually applied to radiomics learning for KRAS mutation detection. He et al.[24] used a residual neural network to estimate the KRAS mutation status from pre-treatment contrast-enhanced CT images of CRC patients, achieving a performance improvement over the prediction model based on radiomics features. Wu et al.[25] presented a model incorporating the handcrafted and deep radiomics features, which can be used for individualized preoperative prediction of KRAS mutations in CRC patients, with a C-index performance of 0.832 for the validation cohort. However, these studies fail to effectively utilize the representational capabilities of deep learning and combine it with hand-crafted radiomics for a more comprehensive exploration of genotype-phenotype correlations.

2.2 Multimodal data fusion

Multimodal data fusion is one of the original topics in multimodal machine learning and has long been investigated by the research community. According to the fusion modes, multimodal data fusion can be categorized into data-level (early) fusion, feature-level (intermediate) fusion, and decision-level (late) fusion. Compared with the data-level fusion and decision-level fusion, feature-level fusion can achieve more flexible and effective data fusion to generate a compact and informative multimodal hidden representation, leading to more variants of fusion methods. The existing feature-level fusion methods mainly include operation-based fusion, tensor-based fusion, subspace-based fusion, and attention-based fusion. The operation-based fusion is to perform simple operations on the feature vectors, such as concatenation, element-wise summation, and element-wise multiplication.

The tensor-based fusion is to conduct outer products across multimodal feature vectors into a higher-order feature matrix to obtain to obtain more powerful feature representation, where the representative works include Tensor Fusion Network (TFN)[26] and Low-rank Multimodal Fusion (LMF)[27]. The subspace-based fusion aims to learn an informative common subspace of multi-modality data, thereby capturing the correlation between different modalities for a more expressive feature representation. Canonical Correlation Analysis (CCA)[28] and its extensions,

Kernel Canonical Correlation Analysis (KCCA)[29] and Deep Canonical Correlation Analysis (DCCA)[30], are the classical subspace-based fusion methods. The attention-based fusion can learn the weights of different modality data via the attention mechanism and incorporate these weighted feature vectors. As the most concerned mode, numerous attention-based fusion methods have been proposed in a wide range of applications, such as Attentional Feature Fusion (AFF)[31], Bilinear Attention Networks (BAN)[32], co-attention[33], and merged attention[34]. However, these studies mostly focus on complex computational mechanisms and neglect the intrinsic characteristics of multimodal data, resulting in suboptimal performance in biomedical applications with limited data amounts.

Recently, a line of work has argued that the key idea of representation learning is to disentangle the substantially lower dimensional and semantically meaningful latent factors from the high-dimensional data, showing that the disentangled representation is beneficial for better representation learning[35, 36]. Disentangled representation learning can take advantage of both the traditional mathematical modeling and machine learning modeling, and has been increasingly applied to multimodal tasks, especially the multimodal sentiment analysis. Hazarika et al.[37] proposed a modality-invariant and modality-specific representation framework to learn factorized subspaces for each modality and provide better representations for multimodal sentiment analysis. Yang et al.[38] presented a feature-disentangled multimodal emotion recognition method to learn the shared and private feature representations for each modality by designing tailored losses for the above subspaces.

3 Method

In this section, we propose DRMF-PaRa for KRAS mutation detection. As shown in Fig. 1, the DRMF-PaRa framework consists of three parts: pathology representation learning, radiology representation

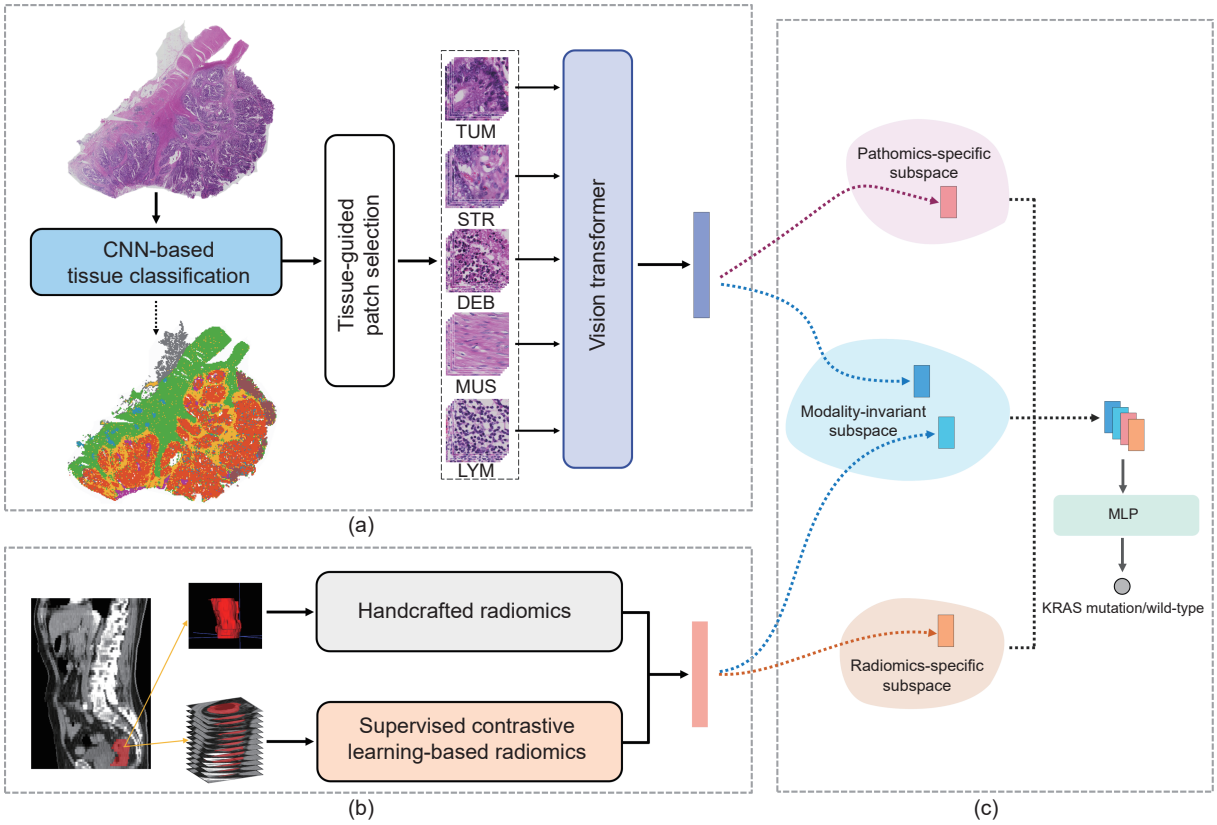


Fig. 1 Overview framework of DRMF-PaRa for KRAS mutation detection. (a) Pathomics learning module, which introduces a tissue-guided Transformer model to extract more comprehensive and targeted pathological features, (b) radiomics learning module, which combines the hand-crafted radiomics and the supervised contrastive learning-based radiomics, and (c) disentangled representation-based multimodal fusion, which learns factorized subspaces for each modality and provides a holistic view for KRAS mutation detection by MultiLayer Perceptron (MLP) from heterogeneous multimodal data.

learning, and multimodal fusion. The details of each part are elaborated as follows.

3.1 Pathomics learning

For a pathology slide with gigapixel resolution, it is common to divide it into thousands of patches, and then randomly select a small number of patches for representation learning. However, this random selection strategy can lead to unstable and underrepresented pathological features. To solve this problem, we propose a tissue-guided Transformer module to capture more comprehensive and targeted pathological features, as shown in Fig. 2. First, a patch-level classifier based on Convolutional Neural Network (CNN) is introduced to classify patches into eight predefined tissue types and the background. Then, dozens of patches are selected from each tissue type to form a representative patch subset of WSIs. Finally, a Transformer network is used to fuse the selected patches to obtain the comprehensive pathological representation.

Tissue classification. To quantify the tissue composition in CRC, we introduce a patch-level tissue classification network. First, pathology slides are tessellated into thousands of patches with a size of 448 pixel×448 pixel and a resolution of 0.25 um/pixel. The deep residual network ResNet-50[39] is used as the patch-level tissue classifier to classify the patches into eight predefined tissue types and the background. The eight tissue types include[40]: adipose tissue (ADI), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal mucosa (NORM), tumor epithelium (TUM), and tumor stroma (STR).

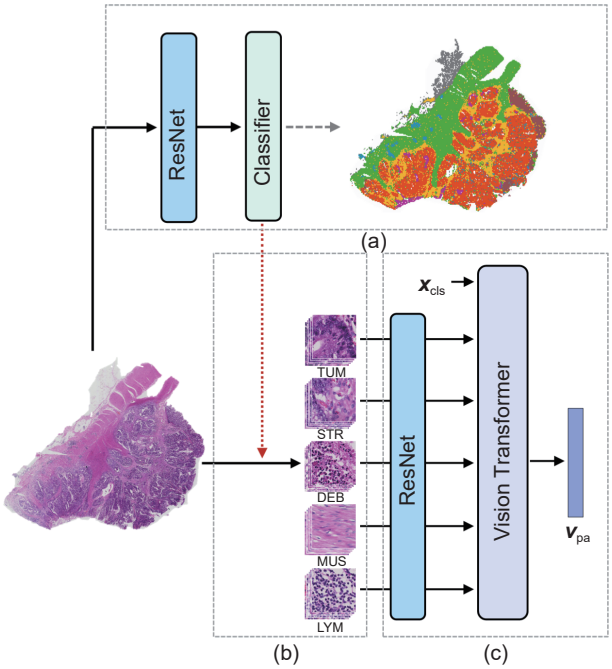


Fig. 2 Tissue-guided Transformer for pathomics learning. (a) CNN-based patch-level tissue classification, (b) tissue-guided patch selection, and (c) patch-level feature fusion based on the Transformer. x_{cls} and v_{pa} represent the learnable class token and the output pathological feature vector, respectively.

Patch selection. Different with random patch selection, we propose a tissue-guided patch selection strategy to obtain more representative patches. A simple approach is to select the same number of patches from each tissue type to form the patch subset. However, this way ignores the proportion of each tissue type in pathology slides and its relevance to tumor disease. Therefore, we can explore the different tissues in patch selection based on tumor relevance, and ultimately select the subsets of tumor epithelium (TUM), stroma (STR), debris (DEB), smooth muscle (MUS), and lymphocytes (LYM) for pathomics learning. Thus, the selected patch set $\boldsymbol{P}$ with a length of $N = 196$ can be formulated as

$$\boldsymbol{P} = \{\boldsymbol{P}_{\mathrm{TUM}}, \boldsymbol{P}_{\mathrm{STR}}, \boldsymbol{P}_{\mathrm{DEB}}, \boldsymbol{P}_{\mathrm{MUS}}, \boldsymbol{P}_{\mathrm{LYM}}\} \tag{1}$$

where $\boldsymbol{P}_{\mathrm{TUM}}$, $\boldsymbol{P}_{\mathrm{STR}}$, $\boldsymbol{P}_{\mathrm{DEB}}$, $\boldsymbol{P}_{\mathrm{MUS}}$, and $\boldsymbol{P}_{\mathrm{LYM}}$ represent the patch subset of tumor epithelium, stroma, debris, smooth muscle, and lymphocytes, respectively.

Transformer-based feature-level fusion. Based on the selected patches, we use the Vision Transformer (ViT-Base)[41], which has a great capability of exploring long-range dependency, to fuse these patches into a comprehensive feature vector. First, the patch $\boldsymbol{p}_i$ is transformed into a 768-dimensional feature vector $\boldsymbol{x}_i$ using the ResNet-50 network in tissue classification. Then, all the feature vectors of the patches in the selected set $\boldsymbol{P}$ can form the feature sequence $\boldsymbol{X}$. Finally, the feature sequence $\boldsymbol{X}$ needs to be concatenated with the learnable class token $\boldsymbol{x}_{\mathrm{cls}}$, and further to be added with the positional embeddings $\boldsymbol{E}_{\mathrm{pos}}$ as the input to the Vision Transformer, which can be formulated as

$$\boldsymbol{X}_0 = \{\boldsymbol{x}_{\mathrm{cls}}, \boldsymbol{X}\} + \boldsymbol{E}_{\mathrm{pos}} \tag{2}$$

where $\boldsymbol{X}_0$ represents the input to the Vision Transformer, and $\boldsymbol{E}_{\mathrm{pos}}$ indicates the one-dimensional sine-cosine positional embeddings to preserve the position information[42].

The process in the ViT can be formulated as

$$\boldsymbol{X}'_k = \mathrm{MSA}\,(\mathrm{LN}\,(\boldsymbol{X}_{k-1})) + \boldsymbol{X}_{k-1},\ k = 1, 2, \ldots, K \quad (3)$$

$$\boldsymbol{X}_k = \mathrm{MLP}\,(\mathrm{LN}\,(\boldsymbol{X}'_k)) + \boldsymbol{X}'_k,\ k = 1, 2, \ldots, K \quad (4)$$

$$\boldsymbol{v}_{\mathrm{pa}} = \mathrm{LN}\,(\boldsymbol{X}_K^0) \quad (5)$$

where $K = 6$ is the number of Transformer encoders, MSA ( ), LN ( ), and MLP ( ) represent the multi-headed self-attention, the layernorm, and the multi-layer perceptron in the Transformer encoder, respectively. As the output of the ViT, the final class token $\boldsymbol{X}_K^0$ has interacted with all feature tokens, and thus can serve as the desired pathological feature representation $\boldsymbol{v}_{\mathrm{pa}}$.

3.2 Radiomics learning

To obtain more effective radiological representation from Computed Tomography (CT) images for KRAS mutation detection, we propose a hybrid radiomics mode that combines the hand-crafted radiomics features and the deep learning based radiomics features. The hand-crafted radiomics features are generic signatures defined by radiology experts based on domain knowledge. The deep learning based radiomics features are task-specific signatures automatically extracted by the deep neural network. Therefore, this hybrid radiomics mode can combine the general and specific features as a comprehensive radiomics representation of CT images. As an initial step, the CT images are labeled with Regions Of Interests (ROIs) by an experienced radiologist.

Hand-crafted radiomics. The hand-crafted radiomics features can be extracted from the ROIs of CT images using the PyRadiomics package[43], which follows the Image Biomarker Standardization Initiative (IBSI)[44]. Specifically, a total of 100-dimensional features were extracted, consisting of three categories: first-order intensity features (18 demensions), shape features (14 demensions), and texture features (68 demensions). First-order intensity features are statistical features based on intensity histograms to characterize the distribution of voxel intensity in the ROIs. Shape features describe the three-dimensional and two-dimensional size characteristics of the ROIs. Texture features capture the correlation relationship between adjacent voxels to describe the structure, heterogeneity, and spatial distribution of the ROIs, which are based on four matrices: (1) Gray Level Co-occurrence Matrix (GLCM), (2) Gray Level Size Zone Matrix (GLSZM), (3) Gray Level Run Length Matrix (GLRLM), and (4) Gray Level Dependence Matrix (GLDM). These hand-crafted radiomics features can provide global information of CT images and have been widely used in radiological representation learning.

Deep learning based radiomics. Although the traditional feature engineering has been proven effective in radiological image analysis, it is difficult to capture the specific signatures associated with KRAS mutation detection. Deep neural networks can automatically extract task-related features, providing a promising direction to solve the above problems. Supervised learning can be used to optimize deep neural networks by minimizing the difference between the predicted outputs and the true labels. The cross-entropy loss is indeed one of the most common objective functions used in supervised learning, especially in classification tasks. Since the optimization of the cross-entropy loss function focuses on the classification properties of high-dimensional features and ignores their distributional properties, it results in a lack of clear semantic information for the high-dimensional feature representation.

Thus, we introduce a Supervised Contrastive learning based Radiomics (SC-Radiomics) method, which implements similarity constraints on high-dimensional features to extract specific radiomics features with clearer semantic information. As shown in Fig. 3, the proposed SC-Radiomics method uses a standard supervised contrastive learning framework, including a feature encoding module and a feature projection module[45]. According to the KRAS status, the CT

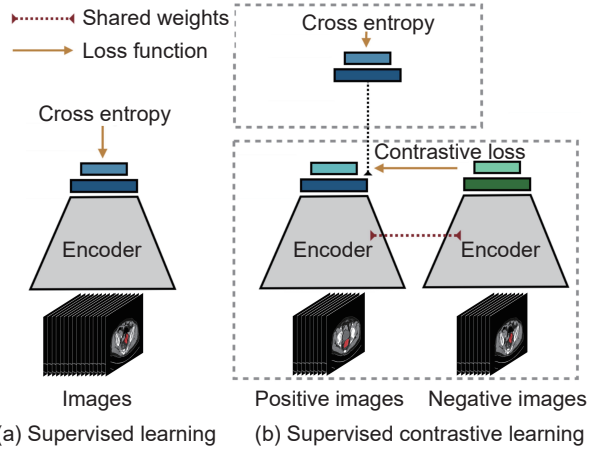


Fig. 3 Deep learning based radiomics learning. (a) Supervised learning based on cross-entropy and (b) supervised contrastive learning.

images are divided into positive sample (mutated) and negative sample (wild-type). Then, the three-dimensional CT ROIs are split into two-dimensional slices, and any combination of three slices can be used as the input to the SC-Radiomics module. In addition to common data augmentation strategies, we adopt both the coarsely labeled mode with bounding box and the precisely labeled mode with region to increase the diversity of samples.

The feature encoding module is based on the backbone of the deep residual network ResNet-50, which maps the input image to a 2048-dimensional representation vector in a unit hypersphere space. The feature projection module consists of two fully connected layers that map the 2048-dimensional representation vector to a 128-dimensional vector. The vector is also normalized to a unit hypersphere space to enable the computation of the contrastive loss. The proposed SC-Radiomics method useing the supervised contrastive loss is defined as following:

$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i=1}^{N} \mathcal{L}_{\mathrm{SCL}}^{i} \tag{6}$$

$$\mathcal{L}_{\mathrm{SCL}}^{i} = \frac{-1}{N_{y_i} - 1} \sum_{j=1}^{N} \left(\mathbb{1}_{i \neq j} \cdot \mathbb{1}_{y_i = y_j} \cdot \log \frac{\exp(f_\theta(\boldsymbol{x}_i) \cdot f_\theta(\boldsymbol{x}_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{i \neq k} \cdot \exp(f_\theta(\boldsymbol{x}_i) \cdot f_\theta(\boldsymbol{x}_k)/\tau)} \right) \tag{7}$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1,2,\ldots,N}$ are the image/label pairs in the sampled minibatch, N_{y_i} is the total number of images with label y_i in the minibatch, f_θ represents the encoder module, $\mathbb{1} \in \{0,1\}$ is an indicator function to judge the condition, and $\tau > 0$ is a scalar temperature parameter. The supervised contrastive loss can encourage closer feature representation among images with the same label to form a more robust representation space. In practice, the representation vector learned by the encoder module can be further mapped to a 668-dimensional deep learning-based radiomics feature vector. Finally, we can combine the hand-crafted radiomics feature vector with the supervised contrastive learning-based radiomics feature vector as the hybrid radiomics feature representation (namely H&SC-Radiomics) $\boldsymbol{v}_{\mathrm{ra}}$.

3.3 Multimodal fusion

Both pathology and radiology images contain richer tumor phenotypic signatures, with the former showing the microscopic features such as structure and morphology, and the latter showing the macroscopic features such as volume and density. On the basis of the pathology representation learning and radiology representation learning, fusing these two phenotypic features can further improve the performance for KRAS mutation detection. The existing studies have generally achieved multimodal feature fusion through sophisticated fusion mechanisms such as tensor-based fusion models and attention-based models. However, these models struggle to address the modality gaps between heterogeneous medical modalities with a small amount of data, resulting in a poor performance.

Therefore, we propose a disentangled representation-based multimodal fusion model, which is inspired by the disentangled representation learning in the multimodal sentiment analysis[37, 38]. Considering that there are both common features and specific features between pathological and radiological representation, we can factorize them into the modality-invariant and modality-specific subspaces to obtain corresponding disentangled feature vectors. These disentangled feature vectors can then be directly combined into a multimodal fusion vector for KRAS gene mutation prediction.

As shown in Fig. 4, the pathological feature vector $\boldsymbol{v}_{\mathrm{pa}}$ and the radiological feature vector $\boldsymbol{v}_{\mathrm{ra}}$ are mapped to a 128-dimensional shared representation $\boldsymbol{h}_{\mathrm{pa}}$ and $\boldsymbol{h}_{\mathrm{ra}}$, respectively, in a common subspace by the modality-invariant encoder. To reduce the discrepancy between the shared representations of each modality, we use the consistency constraint[46] as the similarity loss to encourage them to align together in in the shared subspace, which is defined as

$$\mathcal{L}_{\mathrm{sim}} = \left\| \boldsymbol{h}_{\mathrm{pa}}^{n} \cdot (\boldsymbol{h}_{\mathrm{pa}}^{n})^{\mathrm{T}} - \boldsymbol{h}_{\mathrm{ra}}^{n} \cdot (\boldsymbol{h}_{\mathrm{ra}}^{n})^{\mathrm{T}} \right\|_{\mathrm{F}}^{2} \tag{8}$$

where $\boldsymbol{h}_{\mathrm{pa}}^{n}$ and $\boldsymbol{h}_{\mathrm{ra}}^{n}$ are the normalized feature vectors by the L2-norm, $\|\cdot\|_{\mathrm{F}}^{2}$ represents the squared Frobenius norm.

Meanwhile, the pathological feature vector $\boldsymbol{v}_{\mathrm{pa}}$ and the radiological feature vector $\boldsymbol{v}_{\mathrm{ra}}$ are also factorized into two modality-specific subspaces to generate a 128-dimensional specific representation $\boldsymbol{k}_{\mathrm{pa}}$ and $\boldsymbol{k}_{\mathrm{ra}}$ of each modality, respectively. To ensure that the learned shared representation and the specific representation of each modality focus on the different information, we use the soft orthogonality constraint as the difference loss for these two representation subspaces, which can be formulated as

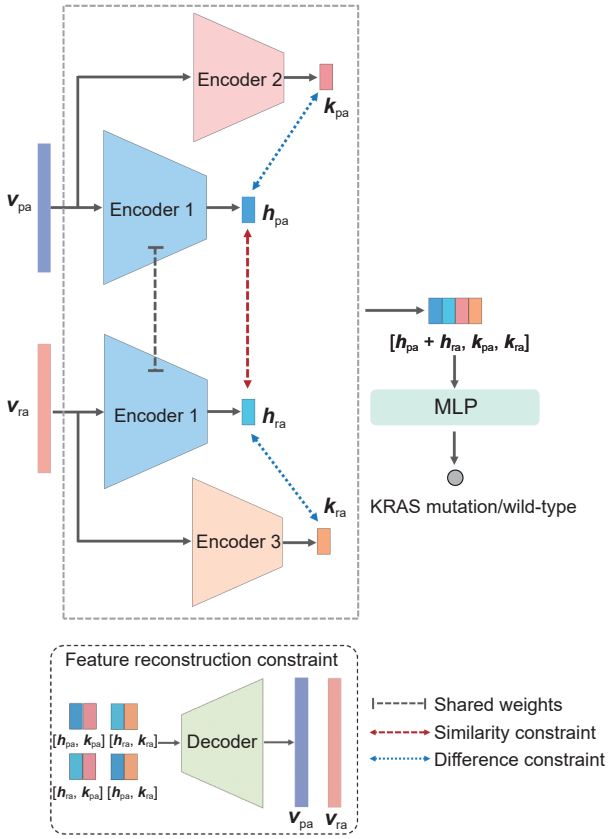


Fig. 4 Disentangled representation-based multimodal fusion model. v_{pa} represents the pathological feature vector, h_{pa} and k_{pa} are the modality-invariant feature vector and the modality-specific feature vector projected from the pathological feature vector. v_{ra} represents the radiological feature vector, h_{ra} and k_{ra} are the modality-invariant feature vector and the modality-specific feature vector projected from the radiological feature vector.

$$\mathcal{L}_{\text{diff}} = \sum_{m\in\{\text{pa, ra}\}} \left\| \boldsymbol{H}_m^{\text{T}} \cdot \boldsymbol{K}_m \right\|_{\text{F}}^2 \tag{9}$$

where $\boldsymbol{H}_m$ is a matrix composed of shared feature vectors $\boldsymbol{h}_m$ in a same batch, $\boldsymbol{K}_m$ is a matrix composed of specific feature vectors $\boldsymbol{k}_m$ in a same batch.

Although the soft orthogonality constraint increases the disparity between the shared representation and the specific representation, it can lead to trivial solutions in the modality-specific subspaces. To ensure the effectiveness of these subspace features, we add a feature reconstruction task that learns the original features from the feature pair of each modality, consisting of the shared features and the specific features. In addition to the feature pairs ($\boldsymbol{h}_{\text{pa}}$, $\boldsymbol{k}_{\text{pa}}$) and ($\boldsymbol{h}_{\text{ra}}$, $\boldsymbol{k}_{\text{ra}}$), we also introduce the cross-modal feature pairs ($\boldsymbol{h}_{\text{ra}}$, $\boldsymbol{k}_{\text{pa}}$) and ($\boldsymbol{h}_{\text{pa}}$, $\boldsymbol{k}_{\text{ra}}$) in the feature reconstruction, which can be formulated as

$$\begin{aligned} \tilde{\boldsymbol{f}}_{\text{pa}} &= D\left(\boldsymbol{h}_m, \boldsymbol{k}_{\text{pa}}; \theta^d\right), \\ \tilde{\boldsymbol{f}}_{\text{ra}} &= D\left(\boldsymbol{h}_m, \boldsymbol{k}_{\text{ra}}; \theta^d\right), \\ m &\in \{\text{pa, ra}\} \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{f}}$ is the reconstructed feature representation, and $D\left(;\theta^d\right)$ represnets the decoder module. The reconstruction task is supervised by the mean squared error loss,

$$\mathcal{L}_{\text{recon}} = \sum \left\| \boldsymbol{f}_m - \tilde{\boldsymbol{f}}_m \right\|_2^2, \quad m \in \{\text{pa, ra}\} \tag{11}$$

where $\|\cdot\|_2^2$ represents the squared L2-norm.

Finally, we can use the explicit combination of the shared vector and the specific vector of two modalities as the multimodal fusion vector $\left[\boldsymbol{h}_{\text{pa}} + \boldsymbol{h}_{\text{ra}}, \boldsymbol{k}_{\text{pa}}, \boldsymbol{k}_{\text{ra}}\right]$ for KRAS mutation detection by an MLP. The total loss function of the disentangled representation-based multimodal fusion model can be formulated as

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{sim}} + \beta\mathcal{L}_{\text{diff}} + \gamma\mathcal{L}_{\text{recon}} \tag{12}$$

where $\mathcal{L}_{\text{BCE}}$ indicates the binary cross-entropy classification loss function, $\alpha = 0.02$, $\beta = 0.03$, and $\gamma = 0.02$ are the weights of the similarity loss, the difference loss, and the reconstruction loss, respectively.

4 Experiment

4.1 Dataset

This study is approved by the Medical Ethics Committee of Beijing Chaoyang Hospital. A total of 111 CRC patients (64 men and 47 women, mean age 64.7 years) who underwent contrast-enhanced CT examination, pathological examination, and KRAS mutation testing between August 2016 and May 2021 are identified retrospectively. There are 56 patients with KRAS mutations and 55 patients with KRAS wild-type, corresponding to 528 pathology slides and 111 sets of contrast-enhanced CT images with slice thickness of 5 mm. They are divided into the training cohort (mutation: 45, wild-type: 45) and the test cohort (mutation: 11, wild-type: 10) with a five-fold cross-validation. In practice, we perform tissue-guided patch selection 200 times for each patient in the training cohort and 20 times for each patient in the test cohort in the pathomics learning. In addition, we use the combination of any three slices from the three-dimensional CT ROIs as the source for deep learning based radiomics learning in the training cohort and the

combination of three consecutive slices in the test cohort.

4.2 Experimental results

KRAS mutation detection can be viewed as a binary classification task that discriminates the mutation type from the wild type. Therefore, we used the accuracy (ACC), sensitivity (SN), specificity (SP), and the Area Under the ROC Curve (AUC) as the metrics to evaluate the overall performance in the KRAS mutation detection.

First, we validate the performance of the proposed DRMF-PaRa framework in KRAS mutation detection based on multimodal data, and compare it with the results of two single-modality representation learning modules. The patch selection strategy is set to TUM (40%), STR (30%), DEB (10%), MUS (10%), and LYM (10%). The experimental results are shown in Table 1, where the pathomics indicates the tissue-guided Transformer in the pathology representation learning module, and radiomics indicates the H&SC-Radiomics in the radiology representation learning module. The corresponding ROC analysis results are shown as Fig. 5. The experimental results show that both the proposed pathomics method and radiomics method achieve good performance in KRAS mutation detection in terms of accuracy, sensitivity, specificity, and AUC. As the multimodal fusion method integrating pathomics and radiomics, the proposed DRMF-PaRa achieves the best performance with accuracy of 0.876, sensitivity of 0.892, specificity of 0.862, and AUC of 0.865.

Table 1 Performance of the proposed DRMF-PaRa framework in KRAS mutation detection.

Method	ACC	SN	SP	AUC
Pathomics	0.857	0.882	0.835	0.851
Radiomics	0.851	0.874	0.829	0.842
DRMF-PaRa	**0.876**	**0.892**	**0.862**	**0.865**

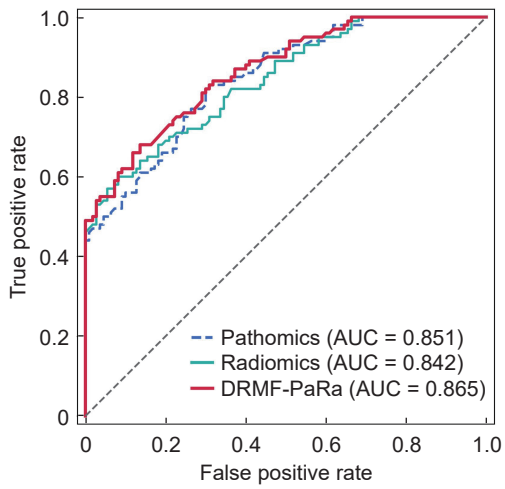


Fig. 5 ROC curve analysis for the proposed DRMF-PaRa framework in KRAS mutation detection.

Then, we compare the proposed DRMF-PaRa with the existing multimodal feature fusion methods, including DCCA[30], TFN[26], LMF[27], AFF[31], BAN[32], Co-Attention[33], and Merged Attention[34]. All the methods use the same pathological feature vector and radiological feature vector as input, which are learned by the tissue-guided Transformer and H&SC-Radiomics. The patch selection strategy is set to TUM (40%), STR (30%), DEB (10%), MUS (10%), and LYM (10%). As shown in Table 2, the experimental results show that the proposed DRMF-PaRa framework outperforms the existing multimodal feature fusion methods in KRAS mutation detection, with accuracy and AUC increased by 1.5% and 1.3%, respectively. This further proves that exploiting the inherent characteristics between multimodal data is more effective than designing the complex fusion mechanisms in specialized fields with limited data volume.

Furthermore, we compare the performance of the proposed single-modality representation learning modules with the existing methods in KRAS mutation detection. The experimental results are shown in Table 3. For pathology images, the proposed tissue-guided Transformer shows the best performance with accuracy of 0.857, sensitivity of 0.882, specificity of 0.835, and AUC of 0.851, outperforming the voting-based models[13, 19] and the multiple instance learning-based models[47, 48]. Meanwhile, compared with the Transformer based on the random patch selection, the

Table 2 Comparison of the proposed DRMF-PaRa framework with different multimodal feature fusion methods in KRAS mutation detection.

Method	ACC	SN	SP	AUC
DCCA[30]	0.859	0.874	0.846	0.845
TFN[26]	0.842	0.878	0.811	0.837
LMF[27]	0.858	0.886	0.833	0.845
AFF[31]	0.849	0.880	0.823	0.837
BAN[32]	0.852	0.886	0.823	0.844
Co-Attention[33]	0.857	0.882	0.835	0.847
Merged Attention[34]	0.861	0.890	0.833	0.852
DRMF-PaRa	**0.876**	**0.892**	**0.862**	**0.865**

Table 3 Comparison of the proposed single-modality representation learning modules with the existing methods in KRAS mutation detection.

Modality	Method	ACC	SN	SP	AUC
Pathology	Coudary et al.[13]	0.816	0.850	0.785	0.801
	Ding et al.[19]	0.794	0.820	0.771	0.782
	Saillard et al.[47]	0.831	0.852	0.812	0.813
	Schirris et al.[48]	0.835	0.852	0.820	0.823
	Vision Transformer	0.852	0.878	0.829	0.843
	Tissue-guided Transformer	**0.857**	**0.882**	**0.835**	**0.851**
Radiology	Yang et al.[14]	0.815	0.852	0.782	0.794
	Taguchi et al.[22]	0.807	0.822	0.795	0.801
	Shiri et al.[23]	0.820	0.852	0.791	0.805
	He et al.[24]	0.816	0.854	0.782	0.798
	Wu et al.[25]	0.825	0.842	0.811	0.815
	SC-Radiomics	0.835	0.852	0.820	0.827
	H&SC-Radiomics	**0.851**	**0.874**	**0.829**	**0.842**

tissue-guided Transformer achieves better performance in KRAS mutation detection, demonstrating the effectiveness of the tissue-guided patch selection strategy. For the radiological images, the proposed H&SC-Radiomics achieves accuracy of 0.851, sensitivity of 0.874, specificity of 0.829, and AUC of 0.842, surpassing the existing hand-crafted radiomics methods. From the experimental results, we can also find that the hand-crafted radiomics and SC-radiomics can capture the different feature information of the radiological images and the combination of the two modes can further improve the performance in KRAS mutation detection.

Finally, we explore the effect of different tissue patch selection in the proposed tissue-guided Transformer on the performance in KRAS mutation detection. The experimental results are shown in Table 4, where the total number of patches for ViT-Base is fixed to 196. The performance of the 100% patches of TUM can be used as the benchmark. First, STR can significantly improve the performance both in accuracy and AUC, with the combination of the 60% tumor epithelium and 40% stroma being the relatively optimal mode. Furthermore, the introduction of DEB, MUS, and LYM can also achieve better performance, but not NORM, MUS, MUC, and ADI. Furthermore, we analyzed the performance of different patch selection of these five tissue types and found that the combination of TUM (40%), STR (30%), DEB (10%), MUS (10%), and LYM (10%) can achieve the best performance. Therefore, we use this patch selection mode as the default mode in the tissue-guided Transformer for the pathological representation learning.

5 Conclusion

Detection of mutations in the KRAS is essential for prognostic analysis and targeted therapy of CRC. Due to long turnaround time and high cost, KRAS mutation detection based on molecular testing cannot be widely used in clinical practice. Recently, the advancement of machine learning, especially deep learning, has greatly promoted the development of KRAS mutation detection from tumor phenotype data such as pathology

Table 4 Performance of the proposed tissue-guided Transformer with different tissue patches in KRAS mutation detection.

Tissue patch	ACC	AUC
TUM (100%)	0.845	0.835
TUM (70%) + STR (30%)	0.849	0.842
TUM (60%) + STR (40%)	0.853	0.844
TUM (50%) + STR (50%)	0.852	0.842
TUM (50%) + STR (40%) + DEB (10%)	0.856	0.844
TUM (50%) + STR (40%) + NORM (10%)	0.850	0.839
TUM (50%) + STR (40%) + MUS (10%)	0.853	0.845
TUM (50%) + STR (40%) + LYM (10%)	0.855	0.845
TUM (50%) + STR (40%) + MUC (10%)	0.851	0.839
TUM (50%) + STR (40%) + ADI (10%)	0.850	0.841
TUM (40%) + STR (40%) + DEB (10%) + MUS (10%)	0.853	0.844
TUM (40%) + STR (40%) + DEB (10%) + LYM (10%)	0.855	0.847
TUM (40%) + STR (40%) + MUS (10%) + LYM (10%)	0.853	0.846
TUM (40%) + STR (30%) + DEB (10%) + MUS (10%) + LYM (10%)	**0.857**	**0.851**
TUM (30%) + STR (40%) + DEB (10%) + MUS (10%) + LYM (10%)	0.855	0.848

slides or radiology images. However, there are still two major problems with existing studies: inadequate single-modal feature learning and lack of multimodal phenotypic feature fusion. In this paper, we propose DRMF-PaRa for KRAS mutation detection. Specifically, the DRMF-PaRa model consists of three parts: (1) the pathomics learning module, which introduces a tissue-guided Transformer model to extract more comprehensive and targeted pathological features; (2) the radiomics learning module, which captures the generic hand-crafted radiomics features and the task-specific deep radiomics features; and (3) the disentangled representation-based multimodal fusion module, which learns factorized subspaces for each modality and provides a holistic view of the two heterogeneous features for KRAS mutation detection. The proposed model was developed and evaluated on a multi-modality dataset of 111 colorectal cancer patients with whole slide images and contrast-enhanced CT. The experiment results demonstrate that the proposed DRMF-PaRa framework can facilitate the KRAS mutation detection in colorectal cancer and shows superiority over the existing methods. In the future, we will apply the DRMF-PaRa framework to more tumor diseases to further promote biomarker detection from multimodal data.

Acknowledgment

The research was supported by the National Natural Science Foundation of China (Nos. 61932018, 32241027, 62072441, 62272326, 62132015, and U22A2037) and the Beijing Municipal Administration of Hospitals Incubating Program (No. PX2021013).

References

[1] S. Alzahrani, H. Al Doghaither, and A. Al-Ghafari, General insight into cancer: An overview of colorectal cancer, *Mol. Clin. Oncol.*, vol. 15, no. 6, p. 271, 2021.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA*: *A Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[3] J. Ferlay, M. Laversanne, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Pi˜neros, A. Znaor, I. Soerjomataram, and F. Bray, Global cancer observatory: Cancer tomorrow, https://gco.iarc.fr/tomorrow, 2022.

[4] W. M. Grady and J. M. Carethers, Genomic and epigenetic instability in colorectal cancer pathogenesis, *Gastroenterology*, vol. 135, no. 4, pp. 1079–1099, 2008.

[5] S. Ogino and A. Goel, Molecular classification and correlates in colorectal cancer, *J. Mol. Diagn.*, vol. 10, no. 1, pp. 13–27, 2008.

[6] D. Senft, M. D. M. Leiserson, E. Ruppin, and Z. A. Ronai, Precision oncology: The road ahead, *Trends Mol. Med.*, vol. 23, no. 10, pp. 874–898, 2017.

[7] C. P. Vaughn, S. D. ZoBell, L. V. Furtado, C. L. Baker, and W. S. Samowitz, Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer, *Genes Chromosom. Cancer*, vol. 50, no. 5, pp. 307–312, 2011.

[8] I. A. Prior, P. D. Lewis, and C. Mattos, A comprehensive survey of ras mutations in cancer, *Cancer Res.*, vol. 72, no. 10, pp. 2457–2467, 2012.

[9] M. Meng, K. Zhong, T. Jiang, Z. Liu, H. Y. Kwan, and T. Su, The Current understanding on the impact of KRAS on colorectal cancer, *Biomed. Pharmacother.*, vol. 140, p. 111717, 2021.

[10] C. De Divitiis, Prognostic and predictive response factors in colorectal cancer patients: Between hope and reality, *World J. Gastroenterol.*, vol. 20, no. 41, p. 15049, 2014.

[11] K. Knickelbein and L. Zhang, Mutant KRAS as a critical determinant of the therapeutic response of colorectal cancer, *Genes Dis.*, vol. 2, no. 1, pp. 4–12, 2015.

[12] C. Tan and X. Du, KRAS mutation testing in metastatic colorectal cancer, *World J. Gastroenterol.*, vol. 18, no. 37, pp. 5171–5180, 2012.

[13] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[14] L. Yang, D. Dong, M. Fang, Y. Zhu, Y. Zang, Z. Liu, H. Zhang, J. Ying, X. Zhao, and J. Tian, Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer? *Eur. Radiol.*, vol. 28, no. 5, pp. 2058–2067, 2018.

[15] O. Reyes, C. Morell, and S. Ventura, Scalable extensions of the Relief F algorithm for weighting and selecting features on the multi-label learning context, *Neuro Computing*, vol. 161, pp. 168–182, 2015.

[16] H.-J. Jang, A. Lee, J. Kang, I. H. Song, and S. H. Lee, Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning, *World J. Gastroenterol.*, vol. 26, no. 40, pp. 6207–6223, 2020.

[17] Y. Jiang, C. K. W. Chan, R. C. K. Chan, X. Wang, N. Wong, K. F. To, S. S. M. Ng, J. Y. W. Lau, and C. C. Y. Poon, Identification of tissue types and gene mutations from histopathology images for advancing colorectal cancer biology, *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 115–123, 2022.

[18] P. L. Schrammen, N. Ghaffari Laleh, A. Echle, D. Truhn, V. Schulz, T. J. Brinker, H. Brenner, J. Chang-Claude, E. Alwers, A. Brobeil, et al., Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology, *J. Pathol.*, vol. 256, no. 1, pp. 50–60, 2022.

[19] K. Ding, M. Zhou, H. Wang, S. Zhang, and D. N.

Metaxas, Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: A retrospective multi-cohort study, *Lancet Digit. Health*, vol. 4, no. 11, pp. e787–e795, 2022.

[20] S. J. Wagner, D. Reisenbüchler, N. P. West, J. M. Niehues, J. Zhu, S. Foersch, G. P. Veldhuizen, P. Quirke, H. I. Grabsch, P. A. van den Brandt, et al., Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study, *Cancer Cell*, vol. 1650, no. 41, pp. 1650–1661.e4, 2023.

[21] M. G. Lubner, N. Stabo, S. J. Lubner, A. M. del Rio, C. Song, R. B. Halberg, and P. J. Pickhardt, CT textural analysis of hepatic metastatic colorectal cancer: Pre-treatment tumor heterogeneity correlates with pathology and clinical outcomes, *Abdom. Imag.*, vol. 40, no. 7, pp. 2331–2337, 2015.

[22] N. Taguchi, S. Oda, Y. Yokota, S. Yamamura, M. Imuta, T. Tsuchigame, Y. Nagayama, M. Kidoh, T. Nakaura, S. Shiraishi, et al., CT texture analysis for the prediction of KRAS mutation status in colorectal cancer via a machine learning approach, *Eur. J. Radiol.*, vol. 118, pp. 38–43, 2019.

[23] R. Shi, W. Chen, B. Yang, J. Qu, Y. Cheng, Z. Zhu, Y. Gao, Q. Wang, Y. Liu, Z. Li, et al., NRAS and BRAF status in colorectal cancer patients with liver metastasis using a deep artificial neural network based on radiomics and semantic features, *Am. J. Cancer Res.*, vol. 10, no. 12, pp. 4513–4526, 2020.

[24] K. He, X. Liu, M. Li, X. Li, H. Yang, and H. Zhang, Noninvasive KRAS mutation estimation in colorectal cancer using a deep learning method based on CT imaging, *BMC Med. Imag.*, vol. 20, no. 1, p. 59, 2020.

[25] X. Wu, Y. Li, X. Chen, Y. Huang, L. He, K. Zhao, X. Huang, W. Zhang, Y. Huang, Y. Li, et al., Deep learning features improve the performance of a radiomics signature for predicting KRAS status in patients with colorectal cancer, *Acad. Radiol.*, vol. 27, no. 11, pp. e254–e262, 2020.

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, arXiv preprint arXiv: 1707.07250.

[27] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in Proc. of *the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Melbourne, Austratia, 2018, pp. 2247–2256.

[28] H. Hotelling, Relations between two sets of variates, *Biometrika*, vol. 28, pp. 321–377, 1935.

[29] P. Lai, Kernel and nonlinear canonical correlation analysis, *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, 2000.

[30] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, Deep canonical correlation analysis, in *Proc. International Conference on Machine Learning*, Atlanta, GA, USA, 2013, pp. 1247–1255.

[31] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, Attentional feature fusion, in *Proc. IEEE Winter Conf. Applications of Computer Vision* (*WACV*), Waikoloa, HI, USA, 2021, pp. 3559–3568.

[32] J.-H. Kim, J. Jun, and B.-T. Zhang, Bilinear attention networks, arXiv preprint arXiv: 1805.07932.

[33] J. Lu, J. Yang, D. Batra, and D. Parikh, Hierarchical question-image co-attention for visual question answering, in *Proc. 30th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 289–297.

[34] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, UNITER: UNiversal image-TExt representation learning, in *Proc. 16th European Conference on Computer Vision*, Glasgow, UK, 2020, pp. 104–120.

[35] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[36] F. Locatello, S. Bauer, M. Lucic, S. Gelly, and O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in *Proc. of the 36th International Conference on Machine Learning,* Long Beach, CA, USA, 2019, pp. 4114–4124.

[37] D. Hazarika, R. Zimmermann, and S. Poria, MISA: Modality-invariant and-specific representations for multimodal sentiment analysis, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 1122–1131.

[38] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, Disentangled representation learning for multimodal emotion recognition, in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 1642–1651.

[39] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 770–778.

[40] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, et al., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, *PLoS Med.*, vol. 16, no. 1, p. e1002730, 2019.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16words: Transformers for image recognition at scale, in *Proc. International Conference on Learning Representations*, arXiv preprint arXiv: 2010.11929.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[43] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C.

Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, Computational radiomics system to decode the radiographic phenotype, *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017.

[44] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, et al., The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping, *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[45] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, Supervised contrastive learning, *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[46] X. Wang, M. Zhu, D. Bo, P. Cui, C. Shi, and J. Pei, AM-GCN: Adaptive multi-channel graph convolutional networks, in *Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, Virtual Event, 2020, pp. 1243–1253.

[47] C. Saillard, O. Dehaene, T. Marchand, O. Moindrot, A. Kamoun, B. Schmauch, and S. Jegou, Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers, in *Proc. MICCAI Workshop on Computational Pathology*, Virtual Event, 2021, pp. 191–205.

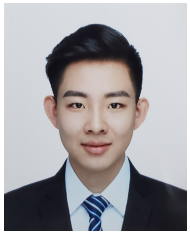
[48] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, and J. Teuwen, DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer, *Med. Image Anal.*, vol. 79, p. 102464, 2022.

Zhilong Lv received the PhD degree in computer science from University of Chinese Academy of Sciences, China in 2023. He is currently a lecturer at School of Computer Science and Technology, Xidian University, China. His current research interests include bioinformatics, data mining, and medical image analysis.

Lin Gao received the PhD degree in circuit and system from Xidian University, China in 2023. She was a visiting scholar at University of Guelph, Canada from 2004 to 2005. She is currently a professor at School of Computer Science and Technology, Xidian University, China. Her research interests include bioinformatics, data mining, graph theory, and optimization.

Ying Wang received the PhD degree in pathology from Capital Medical University, China in 2018. She is the deputy chief physician and associate professor at Department of Pathology Beijing Chaoyang Hospital, Capital Medical University, China. Her current research interests include digestive system diseases, related pathological diagnosis of breast diseases, and biomedical image processing.

Rui Yan received the PhD degree in computer science from University of Chinese Academy of Sciences, China in 2023. He is currently a postdoctoral researcher at School of Biomedical Engineering, University of Science and Technology of China. His research interests include deep learning, bioinformatics, and medical image analysis.



Yuexiao Lin received the Bachelor of Medicine form Capital Medical University, China in 2021, where he is currently a master student. His field of study is clinical surgery. His current research interests include colorectal carcinoma, related clinical diagnosis, surgical treatment, and chemoradiotherapy.

Fa Zhang received the PhD degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences (CAS), China in 2005. He is currently a professor at School of Medical Technology, Beijing Institute of Technology, China. His current research interests include bioinformatics, biomedical image processing, and high-performance computing.