Computers in Biology and Medicine 158 (2023) 106812

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FDTrans: Frequency Domain Transformer Model for predicting subtypes of lung cancer using multimodal data

Meiling Cai [a], Lin Zhao [a], Guojie Hou [a], Yanan Zhang [a], Wei Wu [b], Liye Jia [a], JuanJuan Zhao [a,c], Long Wang [c], Yan Qiang [a,*]

[a] College of Information and Computer, Taiyuan University of Technology, Taiyuan, 030002, China
[b] Department of Clinical Laboratory, Shanxi Provincial People's Hospital, Taiyuan, 030002, China
[c] College of Information, Jinzhong College of Information, Jinzhong, 030002, China

ARTICLE INFO

Keywords:
Deep learning
Histopathological
Lung cancer subtypes
Frequency domain
Multimodal learning

ABSTRACT

Background and purpose: Accurate identification of lung cancer subtypes in medical images is of great significance for the diagnosis and treatment of lung cancer. Despite substantial progress in existing methods, they remain challenging due to limited annotated datasets, large intra-class differences, and high inter-class similarities.

Methods: To address these challenges, we propose a Frequency Domain Transformer Model (FDTrans) to identify patients' lung cancer subtypes using the TCGA lung cancer dataset. We add a pre-processing process to transfer histopathological images to the frequency domain using a block-based discrete cosine transform and design a coordinate Coordinate-Spatial Attention Module (CSAM) to obtain critical detail information by reassigning weights to the location information and channel information of different frequency vectors. Then, a Cross-Domain Transformer Block (CDTB) is designed for Y, Cb, and Cr channel features, capturing the long-term dependencies and global contextual connections between different component features. At the same time, feature extraction is performed on the genomic data to obtain specific features. Finally, the image branch and the gene branch are fused, and the classification result is output through the fully connected layer.

Results: In 10-fold cross-validation, the method achieves an AUC of 93.16% and overall accuracy of 92.33%, which is better than similar current lung cancer subtypes classification detection methods.

Conclusion: This method can help physicians diagnose the subtypes classification of lung cancer in patients and can benefit from both spatial and frequency domain information.

1. Introduction

Lung cancer is one of the most common cancers in the world and is the leading cause of cancer deaths. Non-Small Cell Lung Cancer (NSCLC) accounts for more than 80% of lung cancer cases, while LUng ADenocarcinoma (LUAD) and LUng Squamous Carcinoma (LUSC) are the most common subtypes of NSCLC, with significant differences in treatment and prognosis [1,2]. It is therefore important to distinguish between these two subtypes before starting treatment. Histopathological evaluation of tissue sections by a pathologist is essential in the diagnosis and treatment of lung cancer. However, evaluation by a specialist requires considerable training and expertise and is highly subjective and time-consuming.

In the field of lung cancer diagnosis, deep learning technology has become an integral part of the field. Su et al. [3] proposed a multi-level thresholding image segmentation method based on an enhanced multiverse optimizer to improve the processing efficiency of COVID-19 chest films. Qi et al. [4] presented a new multilevel image segmentation method based on the swarm intelligence algorithm to enhance the image segmentation of COVID-19 X-rays. Hu et al. [5] proposed a short-connection saliency detection network with neutrophil enhancement for polyp region extraction in colonoscopy images. In recent years, many studies have focused on the intelligent diagnosis of lung cancer subtypes. These contents provide the basis for the research of molecular typing. Wang et al. [6] proposed a multi-task deep learning model based on convolutional neural networks using magnified Whole-Slide Images (WSIs) for key cancer lesion region segmentation and histological subtypes classification. Coudray et al. [7] performed a complex classification of histological subtypes. However, the WSIs that are input to the model are too small to be examined by pathologists and thus have little help for doctors to assist in diagnosis. Nair et al.

* Corresponding author.
E-mail address: qiangyan@tyut.edu.cn (Y. Qiang).

https://doi.org/10.1016/j.compbiomed.2023.106812
Received 13 November 2022; Received in revised form 8 March 2023; Accepted 20 March 2023
Available online 22 March 2023

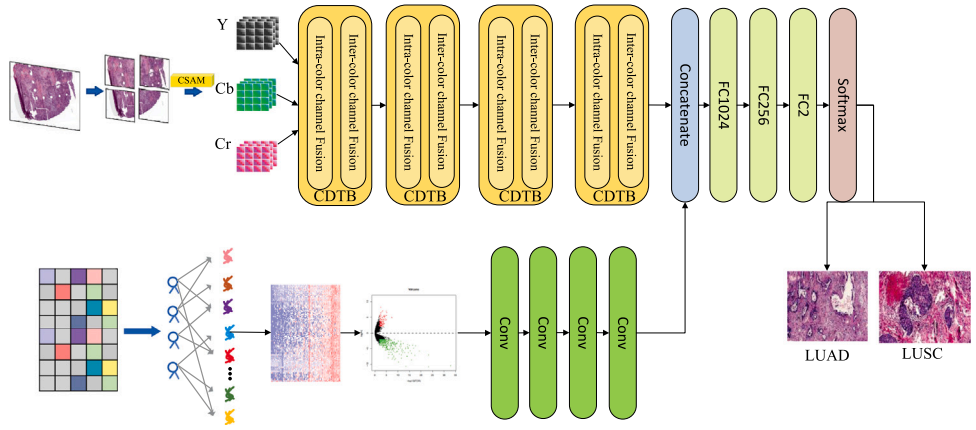

Fig. 1. The proposed FDTrans model architecture. Note that all convolutional layers are followed by Batch Normalization (BN) and Rectified Linear Unit (ReLU) (not included in the figure).

[8] investigated whether the Inception V3 convolutional neural network could discriminate lung cancer subtypes from ambiguous WSIs, and demonstrated that datasets with different resolutions affect model prediction accuracy.

Due to the gigapixel size of WSIs and the limitations of GPU memory, it is not feasible to train the network in an end-to-end manner with WSIs. Therefore, many approaches utilize Multiple Instance Learning (MIL) to train patch classifiers from patches extracted from WSIs and then aggregate the results to the WSI level. Zhao et al. [9] developed a multi-resolution expectation–maximization convolutional neural network method based on MIL to localize ROIs, and in the case of only WSI-wise labels, the network was selected to recognize image patches to train the network for subsequent classification. Hashimoto et al. [10] developed a novel CNN-based method for cancer subtypes classification by combining multiple instances, domain adversarial, and multi-scale learning frameworks. Li et al. [11] introduced a MIL aggregator to model the relationship between instances in a two-stream architecture with a trainable distance metric for classification tasks. Although this method is effective, it is computationally expensive because it takes time to tile the WSI into small pieces, and the spatial relationships between these tiles cannot be explored.

Recently, inspired by the success of Transformer [12] in various NLP tasks, more and more Transformer-based methods have appeared in CV tasks. The Transformer splits the image into multiple patches and provides a sequence of linear embeddings of these patches as input to the Transformer, establishing global connections between the sequence's tokens. The ability to model long-range dependencies is also applicable to medical image-based tasks. TransUNet [13] utilized CNN to extract features and then feed them into Transformer for long-term dependency modeling. DS-TransUNet [14] adopt a dual-scale encoder sub-network based on Swin Transformer to extract coarse-grained and fine-grained feature representations at different semantic scales to improve the quality of semantic segmentation of different medical images. The proposed TransMIL [15] explored both morphological and spatial information and can effectively handle imbalanced/balanced and binary/multivariate classification with good visualization and interpretability.

Aiming at the problem that image downsizing will lead to information loss and accuracy degradation, some works [16,17] reduce the information loss by learning task-aware downscaling networks. However, these networks are task-specific and require additional computation. Some works focus on the frequency domain and use it for image-understanding tasks. In [18,19], the frequency was introduced in CNN by JPEG encoding. Xu et al. [20] reconstructed high-resolution images in the frequency domain, and then fed the reconstructed Discrete Cosine Transform (DCT) coefficients to the CNN model for inference. Qin et al. [21] treats the channel representation problem as a compression

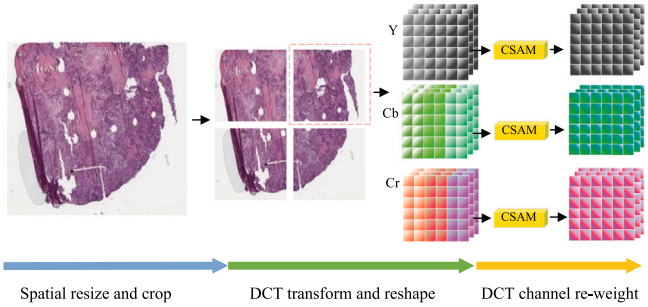

Fig. 2. Data pre-processing for converting histopathological images to frequency domain.

process using frequency analysis, and proposes a multi-spectral channel attention method. Ma et al. [22] combined spatial domain information and frequency domain information to determine KRAS mutation status in colorectal cancer patients.

Due to the complexity of natural phenomena, a single modal feature is not sufficient to provide a complete understanding of the analysis, so for complex analysis, multimodal data is more advantageous [23]. Therefore, more and more methods choose to fuse multimodal data for analysis. Liu et al. [24] combined the patient's genetic modality data with the image modality data, established feature extraction networks according to different morphologies and states, and finally used the fusion features to predict breast cancer subtypes. Chen et al. [25] fused tissue images and genomic features for end-to-end training to predict survival outcomes in lung cancer patients. Braman et al. [26] combined information from multiparametric MRI examinations, biopsy-based modalities (such as H&E slice images and/or DNA sequencing), and clinical variables into a comprehensive multimodal risk score to complete the diagnosis of glioma patients. Zhang et al. [27] combined image features and patient differential genes, designed an ant colony algorithm based on Maximum Information Coefficient Correlation (MICC-ACO) for unsupervised feature selection of fused features, which was finally used for patient lung cancer subtypes classification.

Although the above methods can achieve good results, there are still some limitations, namely: (1) in order to meet the input requirements of the classification network, a large number of real images are scaled down on a large scale, and the image reduction will inevitably lead to information loss and loss of precision; (2) due to the characteristic that deep networks prefer low-frequency information, some frequency channels containing high-level detailed information will be sacrificed to improve the overall model performance; (3) most of the developed and validated relevant features are directly extracted from the spatial

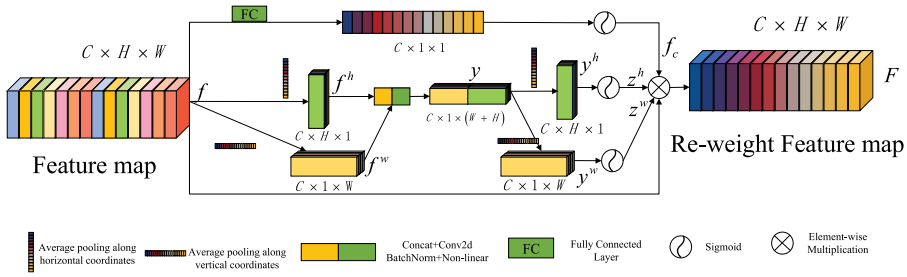

Fig. 3. The architecture of Coordinate-Spatial Attention Module (CSAM).

domain (shape and texture, etc.) and cannot adequately distinguish lung cancer subtypes.

To address the above issues, this paper proposes a Frequency Domain Transformer model (FDTrans) for discriminating LUSC and LUAD from histopathological images and genomic data of lung cancer patients. First, the WSI is converted to the YCbCr color space, and the block-based 2D-DCT is used to obtain the spatial spectrum; secondly, we design a Coordinate-Spatial Attention Module (CSAM) for the spatial spectrum of each color channel, redistribute the weight coefficients of the low-frequency information channel and the high-frequency information channel, and obtain full-frequency information with orientation awareness and position sensitivity. Then, a Cross-Domain Transformer Block (CDTB) is designed to fuse features from different color channels, capturing long-term dependencies and global contextual connections between different features. Finally, the image and gene branches are fused and fed into a fully connected layer to output the classification result.

The main contributions of our work can be summarized in four aspects:

- To the best of our knowledge, this is the first work that combines frequency and spatial domains to analyze histopathological images of lung cancer patients to determine lung cancer subtypes, as shown in Fig. 1.
- A Frequency Domain Transformer model (FDTrans) is proposed, which converts the original histopathological images to the frequency domain through lossless compression, and then combines the gene information to obtain the classification results of lung cancer subtypes, as shown in Fig. 2.
- The Coordinate-Spatial Attention Module (CSAM) is designed, which can obtain key details by automatically adjusting the position information and channel information weight of full-frequency information, as shown in Fig. 3.
- The Cross-Domain Transformer Block (CDTB) is constructed to enhance the fusion of different color space vectors by cross-attention, combining convolutional projection to reduce information loss and fully extract spatial information, capturing long-term dependencies and global contextual connections between different features, as shown in Fig. 4.

2. Related work

In this section, we first summarize methods related to frequency domain applications to images, and then provide an overview of recent related work on vision transformers, especially for medical images.

A. Frequency Domain

In fact, frequency analysis is a common and important method in digital image processing and has been widely used in various tasks in computer vision. Most of them use Discrete Fourier Transform (DFT), Wavelet Transform (WT) or Discrete Cosine Transform (DCT) to transform the spatial image into the frequency domain. Zhong et al. [28] introduce the frequency domain as an additional cue to better detect camouflaged objects from the background. F3-Net [29] uses DCT to extract frequency-domain information and analyzes statistical features for face forgery detection. Liu et al. [30] combines spatial images and phase spectra to capture face forgery upsampling artifacts to improve face forgery Transferability of forgery detection.

B. Vision Transformer

Recently, Transformers have shown a strong ability to extract global information, which makes up for the shortcomings of CNN, and more and more Transformer-based methods appear in CV tasks. TransFuse [31] is proposed to improve the efficiency of global context modeling by fusing transformers and neural networks. Furthermore, to efficiently train the model on medical images, MedT [32] introduces the gated Axial Attention [33] based on the axial depth lab. Also, transformers are not sensitive to details. Therefore, some methods combining CNNs with transformers have been proposed. For example, Transunet [13] and Transbts [34] apply convolution and transformation to extract low-level features and global information, respectively. They effectively reduce the number of parameters.

3. Methods

3.1. Overview

In this section, the overall structure of the proposed Frequency Domain Transformer model (FDTrans) is detailed, as shown in Fig. 1. First, we convert the WSI to the YCbCr color space, obtain the spatial spectrum through DCT, and use the Coordinate-Spatial Attention Module(CSAM) to automatically adjust the weight coefficient between the low-frequency information channel and the high-frequency information channel, capturing long-range dependencies and preserving accuracy position information to obtain full-frequency information with orientation awareness and position sensitivity. Since the luminance component Y and the chrominance components Cb and Cr have different resolutions, we choose to input the feature maps of the Y, Cb, and Cr channels into the Cross-Domain Transformer Block (CDTB), and use cross attention to strengthen the fusion of different color component features., which captures long-term dependencies and global context connections between different features. Finally, gene features with temporal and spatial specificity are added, and fusion analysis with image features is used to determine lung cancer subtypes. In the following sections, we detail the pre-processing method in FDTrans to convert WSI to the frequency domain and then elaborate on the CSAM for redistributing the channel weights of different frequency components. Finally, we show that our FDTrans can benefit from the YCbCr component decentralized processing design, and describe how CDTB can effectively fuse YCbCr component features.

3.2. Pre-processing

The pre-processing process of converting WSI to frequency domain is shown in Fig. 2. We transform the image into YCbCr color space, divide it into 8 × 8 blocks according to the JPEG compression standard [35] in the luminance component Y and chrominance components

Cb, Cr, and calculate the two-dimensional forward DCT of each block. In general, the function of the two-dimensional DCT is:

$$f(h,w) = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j} cos\left(\frac{h\pi}{H}\left(i+\frac{1}{2}\right)\right) cos\left(\frac{w\pi}{W}\left(j+\frac{1}{2}\right)\right) \quad (1)$$

where, $f(h,w)$ is the coefficient after DCT transformation, $x_{i,j}$ is the original input, H is the height of $x_{i,j}$, and W is the width of $x_{i,j}$.

We combine all 8×8 blocks into one channel, keeping their spatial relationship at each frequency, grouping 2D DCT coefficients of the same frequency into one channel to form a 3D DCT cube. Thus, each of the Y, Cb, and Cr channels provides $8 \times 8 = 64$ channels. The original RGB input image is assumed to have the shape $H \times W \times C$, where $C = 3$, and the height and width of the image are denoted as H and W respectively. After converting to the frequency domain, the feature shapes of the Y, Cb, and Cr channels become $H/8 \times W/8 \times 64$. After obtaining the features of the Y, Cb, and Cr channels, we use CSAM to assign different weights to the different channels to maximize the low-frequency information that represents the contours of the function, while retaining the other high-frequency components that determine the detailed information.

3.3. Coordinate-Spatial Attention Module(CSAM)

In our model, the different channels of the feature vectors of the Y, Cb, and Cr channels are at different frequencies and, when analyzed according to the characteristics of the frequencies, almost all frequency components have a small gap between the lowest frequency components [21], indicating that the other frequency components also cope well with the subsequent task.

Furthermore, convolution can only capture local relationships, but cannot model long-term dependencies that are critical to the visual task. Therefore, we propose a frequency coordinate attention mechanism that preserves precise location information to exploit the relative importance of location information to different frequency channels. We combine channel attention with coordinate attention, and use multi-spectral channel attention to replace the traditional GAP method in channel attention, and learn different weights for all frequency channel components. Fig. 3 describes the calculation process of CSAM, which can be expressed as

$$F = f \times f_c \times z^h \times z^w \quad (2)$$

where, f is the input feature, f_c is the weight generated by the multispectral channel attention, and z^h, z^w are the attention weights in the horizontal and vertical directions generated by the coordinate attention. F is the output feature of CSAM. The details of multispectral channel attention and coordinate attention are presented below.

Multispectral Channel Attention. To efficiently compress the global spatial information into channel attention, most attention mechanisms use scalars to represent channels, which in turn leads to a large amount of information loss. Since our input feature has the characteristic that one channel is at one frequency, we will use a scalar to encode the information of one frequency channel, which can effectively utilize the representation ability and correlation of different frequency channels. Specifically, the input f represents the frequency channel, $\left[f_0, f_1, \ldots f_{n-1}\right]$ which represents the channels of different frequencies that make up f, the whole multispectral channel attention can be written as:

$$f_c = Sigmoid(FC(f)) \quad (3)$$

where, $Sigmoid(\cdot)$ is activation function, $FC(\cdot)$ is fully connection. Only the lowest frequency GAP is retained for comparison, which can effectively enrich the compressed full-frequency channel information.

Coordinate attention. To implement attention blocks that capture remote interactions spatially through precise location information, we aggregate features along two spatial directions separately, one capturing remote dependencies and one retaining precise location information, resulting in a pair of direction-aware feature maps. Specifically, given an input f, we encode each channel along the horizontal and vertical coordinates using two spatially scoped pooling kernels to obtain a pair of one-dimensional features f^h, f^w. After connecting them and sending them to a shared 1×1 convolution activation, they are again divided into two separate tensors along the spatial dimension, which are each convolutionally transformed into a tensor with the same number of channels as the input f. The activation gives the attention weights. Thus, the whole coordinate attention can be expressed as:

$$\begin{aligned} &f \rightarrow f^h, f^w \\ &y = Sigmoid\left[W_1 Concat(f^h, f^w)\right] \\ &y \rightarrow y^h, y^w \\ &z^h = Sigmoid(W_h y^h), z^w = Sigmoid(W_w y^w) \end{aligned} \quad (4)$$

where, $Sigmoid(\cdot)$ is activation function, $W_i, i = 1, w, h$ is 1×1 convolution, $Concat(\cdot)$ is concatenation function, y is the intermediate vector, y^h and y^w are one-dimensional features at the horizontal and vertical levels, and the above transformations, unlike the channel attention that generates individual feature vectors, help the network to generate spatially selective attention maps more accurately.

3.4. Cross Domain Transformer Block (CDTB)

For pathological images, our goal is to enhance their unique color information. We transfer pathological images into YCbCr space and separate luminance and chrominance information into Y, Cb, and Cr channels. The designed CDTB framework is shown in Fig. 4. Since the location information is very important relative to the diagnosis of lesions, we use depthwise convolution to embed features for the input three-channel feature maps instead of linear projections of the original transformer. Compared with linear projection, convolution projection can preserve the spatial location information of pixels and does not require location embedding. Then, global interactions within the same domain are efficiently integrated using intra-color channel fusion. The basic component of intra-color channel fusion is multi-head self-attention, which can establish the correlation between distance pixels and capture information from different angles. Second, convolution is used to optimize the features generated by MSA after splicing, and convolution projection can reduce the loss of information and fully extract spatial information. Finally, residual connections are used for MSA and Conv. The intra-color channel fusion can be expressed as follows (taking the Y color channel as an example):

$$\left\{Q_1^Y, K_1^Y, V_1^Y\right\} = \left\{F^Y W_1^Q, F^Y W_1^K, F^Y W_1^V\right\} \quad (5)$$

$$\widehat{z_1^Y} = \mathrm{MSA}\left(Q_1^Y, K_1^Y, V_1^Y\right) + F^Y \quad (6)$$

$$z_1^Y = Conv\left(\widehat{z_1^Y}\right) + \widehat{z_1^Y} \quad (7)$$

where, F^Y represents the input feature map of the Y color channel, $W_i^m, m = Q, K, V$ denotes depthwise convolution operation, $MSA(\cdot)$ indicates multi-head self-attention, $\widehat{z_1^Y}$ represents the output of multi-head self-attention, z_1^Y indicates the output of the Y color channel in the intra-color channel fusion, $Conv(\cdot)$ denotes 3×3 convolution operation.

After intra-color channel fusion, we design inter-color channel fusion, which further integrates the global interactions between different color channels. Both intra-color channel fusion and inter-color channel fusion follow similar baselines. The main difference is that the inter-color channel fusion adopts Multi-Head Cross-Attention (MCA) to achieve global context exchange across color channels. We denote the feature of the current color channel as Q, and the features of the other two color channels as K and V, calculate the similarity between Q, K, and V through MCA, merging the information across the color channels

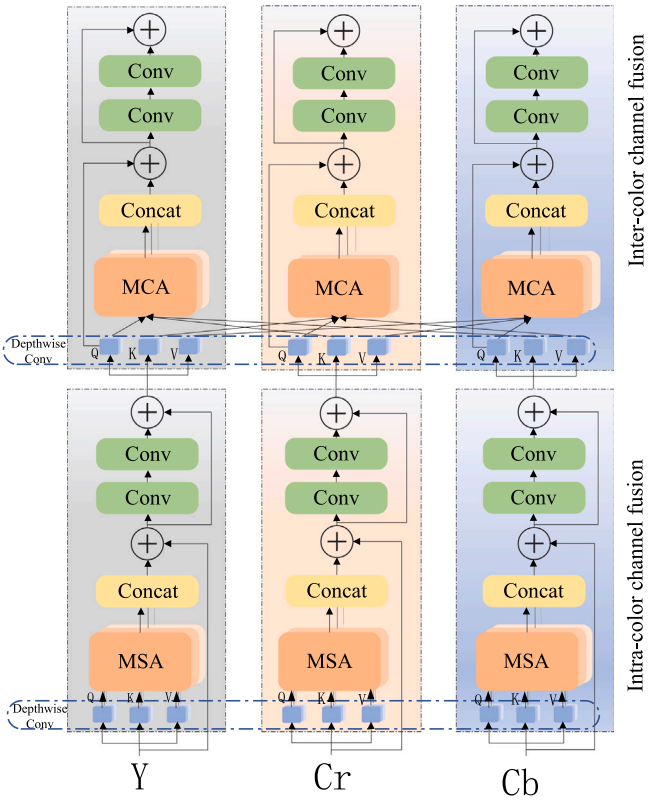

Fig. 4. The architecture of Cross Domain Transformer Block (CDTB). Note that the CDTB consists of two stages: intra-color channel fusion and inter-color channel fusion.

while retaining the information of the current color channel through residual concatenation. The inter-color channel fusion can be expressed as follows (using the Y color channel as an example):

$$\{Q_2^Y, K_2^Y, V_2^Y\} = \left\{z_1^Y W_2^Q, z_1^Y W_2^K, z_1^Y W_2^V\right\} \tag{8}$$

$$\widehat{z_2^Y} = \mathrm{MCA}\left(Q_2^Y, K_2^{C_b}, V_2^{C_r}\right) + Q_2^Y \tag{9}$$

$$z_2^Y = Conv\left(\widehat{z_2^Y}\right) + \widehat{z_2^Y} \tag{10}$$

where, $\mathrm{MCA}(\cdot)$ indicates multi-head cross-attention, $\widehat{z_2^Y}$ represents the output of multi-head cross-attention, z_2^Y indicates the output of the Y color channel in the inter-color channel fusion.

4. Data

The TCGA-NSCLC (The Cancer Genome Atlas - Non-Small Cell Lung Cancer) dataset includes two subtypes of lung cancer, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). The dataset contains a total of 1053 diagnostic WSIs. 1009 patients are retained after expert screening for patients with both histopathological image data and genetic data. All data are randomly divided into a training dataset and a test dataset in a ratio of 8:2. Regarding the processing of gene expression information, we first screened the original gene expression profile for differential genes, and eliminated genes with similar expression levels in the two different lung cancer subtypes. The screening condition is that the difference between genes is more than double and the corrected p-value value must be less than 0.01.

5. Experiments and results

5.1. Implementation details

Our experiments are mainly done on NVIDIA TITAN Xp GPU. All models involved in the experiments were trained with 10-fold cross-validation, and the number of epochs per fold was set to 20. Adam is used as our optimizer. The initial learning rate is set to 0.001. The batch size is set to 1.

5.2. Evaluation metrics

To quantitatively evaluate the proposed FDTrans, we used ACcuracy (AC), SEnsitivity (SE), SPecificity (SP), F-measure and AUC as performance metrics to assess the classification results obtained and used AUC to rank the classification performance of each model on the test dataset.

5.3. Ablation experiment

In this section, we evaluate the impact of CSAM, CDTB, and color channel multipath architectures on the overall FDTrans model performance, respectively.

5.3.1. Ablation study of CSAM

We compare several attention methods with CSAM to objectively evaluate the performance of our CSAM. The experimental results are shown in Table 1. Compared with other attention methods Fca [21], CBAM [36], SE [37], and CA [38], CSAM achieves the best results in terms of AUC, AC, and SP, although it is slightly lower than Fca in terms of SE, which shows that our CSAM is meaningful and can obtain full-frequency information with orientation awareness and position sensitivity, and improve the overall FDTrans model performance. Especially compared with other models, our CSAM has achieved the highest value on SP. Our model can effectively use the information of each frequency to solve the problem of high clinical misdiagnosis rate, and it is easier to detect non-disease patients and improve the predictive accuracy of the model.

5.3.2. Ablation study of CDTB

Then, we evaluate the effect of CDTB on the classification performance of FDTrans, as shown in Table 2. We replace the CDTB of FDTrans with a convolutional block. Furthermore, we also compare CDTB and STB (Swin Transformer Block [39]). Compared with using convolution blocks and using STB blocks, FDTrans with CDTB is significantly improved. Demonstrate the importance of CDTB that fuses the advantages of convolution and multi-head cross-attention in the Transformer. It establishes correlations between features across color channels while maintaining detailed information. In addition, CDTB effectively fuses image information and specific genetic information, which is more beneficial compared with other modules to assist the model in clinically detecting the misdiagnosis rate in the diseased population and taking more appropriate treatment in time for two different subtypes of lung cancer.

5.3.3. Ablation study of color channel multipath architecture

Finally, we test the effect of the color channel multipath architecture on the classification performance of FDTrans. Seen in Table 3. The baseline network adopts a single-path architecture. We concatenate the features of the three color channels as input. Compared with baseline networks, FDTrans achieves the best results on all four metrics. The experimental results show that TranSiam can more fully utilize the complex visual features of multi-color components. For histopathological images, the color information they contain is very important, and clinically, in order to facilitate observation under the microscope and make an accurate pathological diagnosis, pathologists need to stain sections in various colors. The experimental results show that FDTrans can more fully utilize the composite frequency features of multiple color components.

Table 1
Comparison of CSAM with other attention methods on the validation set. We give the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) for each model for the entire lung cancer subtypes classification. Without indicates no means and no attention methods are used.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
Without	90.97 ± 0.02	89.29 ± 1.89	84.05 ± 2.01	91.26 ± 2.10	92.52 ± 1.38
Fca	92.78 ± 0.56	91.93 ± 0.40	87.43 ± 1.78	**93.63 ± 0.56**	94.40 ± 0.27
CBAM	91.64 ± 0.85	90.78 ± 1.02	86.11 ± 1.67	92.54 ± 1.23	93.58 ± 0.74
SE	91.17 ± 0.34	90.65 ± 0.24	84.90 ± 1.85	92.81 ± 0.64	93.51 ± 0.17
Coordinate attention	92.16 ± 0.52	91.31 ± 0.80	86.59 ± 1.91	93.08 ± 0.83	93.96 ± 0.56
CSAM	**93.16 ± 1.07**	**92.33 ± 0.48**	**90.09 ± 3.01**	93.17 ± 0.89	**94.64 ± 0.32**

Table 2
Comparison of CDTB with other methods on the validation set. We present the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) of each model across the entire lung cancer subtypes classification. w/o means without, w/ means with.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
FDTrans w/o CDBT	88.66 ± 0.74	87.64 ± 1.82	82.24 ± 1.66	89.67 ± 2.25	91.32 ± 1.36
FDTrans w/ SBT	89.89 ± 0.56	88.70 ± 2.23	83.33 ± 2.17	90.72 ± 2.42	92.09 ± 1.63
FDTrans w/ CDBT	**93.16 ± 1.07**	**92.33 ± 0.48**	**90.09 ± 3.01**	**93.17 ± 0.89**	**94.64 ± 0.32**

Table 3
Comparison of one-path and three-path on the validation set. We present the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) of each model across the entire lung cancer subtypes classification.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
Single path	88.13 ± 0.85	87.44 ± 2.29	82.24 ± 1.88	89.40 ± 2.52	91.17 ± 1.69
Triple path	**93.16 ± 1.07**	**92.33 ± 0.48**	**90.09 ± 3.01**	**93.17 ± 0.89**	**94.64 ± 0.32**

Table 4
Comparison of FDTrans with five other medical image classification methods on the validation set. We present the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) of each model across the entire lung cancer subtypes classification.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
Swin Transformer	84.78 ± 1.85	83.01 ± 1.18	76.81 ± 3.15	85.35 ± 0.55	87.95 ± 0.79
DSMIL	87.53 ± 0.64	86.62 ± 1.85	83.09 ± 3.44	87.94 ± 1.38	90.52 ± 1.32
Local-learning WSI	91.78 ± 1.43	90.41 ± 1.40	**90.57 ± 2.96**	90.35 ± 1.45	93.19 ± 1.01
MMDL	83.77 ± 0.76	82.75 ± 0.64	76.44 ± 2.20	85.12 ± 0.40	87.76 ± 0.42
MICC-ACO	82.59 ± 2.01	81.99 ± 0.86	75.36 ± 2.51	84.49 ± 0.82	87.21 ± 0.60
FDTrans	**93.16 ± 1.07**	**92.33 ± 0.48**	90.09 ± 3.01	**93.17 ± 0.89**	**94.64 ± 0.32**

5.4. Comparison experiment

Here, we compare FDTrans with popular medical image classification methods. On the TCGA lung cancer dataset, FDTrans is compared with popular medical image classification methods, including Swin Transformer [39], DSMIL [11], Local-learning WSI [40], MMDL [41], MICC-ACO [27]. We reproduce the above method under the same benchmark, and the comparison results are shown in Table 4. Our FDTrans outperforms five other medical image classification models on key evaluation metrics. Compared with Local-learning WSI, our FDTrans improves AUC by 1.38%, and Local-learning performs the second best. Although our FDTrans is slightly lower than Local-learning WSI in SP value, it is 1.92% and 2.82% higher in AC and SE value, which are important metrics for evaluating classification performance. These results all demonstrate that our FDTrans can capture long-term dependencies and global contextual connections between different features, leading to better lung cancer subtypes classification results. We also plot the AUC curves of our FDTrans and five other models in Fig. 5 to more intuitively show the classification performance of our FDTrans.

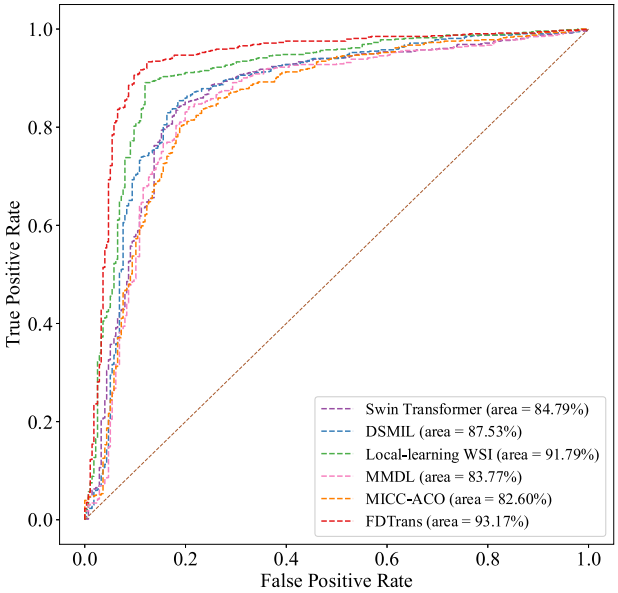

Fig. 5. ROC plots of our FDTrans and five other models.

6. Discussion

6.1. The need for preprocessing

In our FDTrans, we propose a pre-processing process to convert the original histopathological image to YCbCr color space, obtain the spatial spectrum by DCT, and obtain uncompressed full-frequency information after weight redistribution for each channel.

To verify the necessity of pre-processing, we compare in Table 5 the effect of using and without pre-processing on the classification performance of the FDTrans model. As shown in Table 5, the preprocessing process can losslessly compress histopathological images, and save the detail information and contour information according to different weights, which has a very significant effect on the performance improvement of our model, and the full frequency information is of great help to the classification of lung cancer subtypes.

Table 5
Comparison on the validation set with and without pre-processing. We present the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) of each model across the entire lung cancer subtypes classification. w/o means without, w/ means with.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
FDTrans w/o pre-process	85.32 ± 0.45	84.01 ± 0.20	78.14 ± 1.46	86.22 ± 0.40	88.68 ± 0.12
FDTrans w/ pre-process	**93.16 ± 1.07**	**92.33 ± 0.48**	**90.09 ± 3.01**	**93.17 ± 0.89**	**94.64 ± 0.32**

Table 6
Comparison on the validation set with and without gene expression data. We present the AUC (%), AC (%), SP (%), SE (%) and F-measure (%) of each model across the entire lung cancer subtypes classification. w/o means without, w/ means with.

Methods	AUC (%)	AC (%)	SP (%)	SE (%)	F-measure (%)
FDTrans w/o gene	85.63 ± 0.59	85.39 ± 0.49	83.57 ± 0.55	86.08 ± 0.89	89.54 ± 0.41
FDTrans w/ gene	**93.16 ± 1.07**	**92.33 ± 0.48**	**90.09 ± 3.01**	**93.17 ± 0.89**	**94.64 ± 0.32**

6.2. The advantages of multimodality

To discuss the effectiveness of multimodal data, we compare the classification performance of the FDTrans model with and without gene expression data in Table 6. As shown in Table 6, the AUC of the model using both histopathological image features and gene features is significantly higher than that of the model using only single modality features. Since our model adds genetic information, which can make up for the specific information that is ignored in the single-modal feature extraction of images. This can effectively reduce the misdiagnosis rate and improving the sensitivity of the model, thereby achieving better lung cancer subtype classification results.

7. Conclusion and future work

In summary, we propose a dual-branch deep learning model that combines image domain and genetic information to determine lung cancer subtypes in patients. In the gene branch, we employ machine learning methods to extract specific features. In the image branch, we obtain the spatial spectrum through DCT, and input it into the designed CSAM, after redistributing the weights to each channel, we obtain the full-frequency information features of the three color components without compression. Following this, we design a CDTB to enhance the fusion of different color space vectors through multi-head cross-attention, capturing long-term dependencies and global contextual connections between different features. Finally, we connect the two branches, fuse and output the classification result. The experimental results show that the method extracts image features from spatial and frequency domains, utilizes gene-specific features, and has better classification performance than existing classification methods.

Although satisfactory predictive results have been achieved in the classification of lung cancer subtypes, our model has some limitations. First, we only validated on a single public dataset TCGA. In addition, our study did not use CT images, but other studies have shown that CT images, as a noninvasive method, can also be used to predict lung cancer subtypes.

Future research can be done in two directions: (1) We plan to extend the proposed model to a semi-supervised learning framework, using unlabeled cases as training samples. (2) Other imaging information of CRC patients needs to be integrated into the deep learning model to obtain more accurate classification results.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61872261); the National Natural Science Foundation of China (Grant No. U21A20469); the National Natural Science Foundation of China (Grant No. 61972274); and the Natural Science Foundation of Shanxi Province, China (Grant No. 202103021224066).

References

[1] Y. Ma, W. Feng, Z. Wu, M. Liu, F. Zhang, Z. Liang, C. Cui, J. Huang, X. Li, X. Guo, Intra-tumoural heterogeneity characterization through texture and colour analysis for differentiation of non-small cell lung carcinoma subtypes, Phys. Med. Biol. 63 (16) (2018) 165018.

[2] X. Zhu, D. Dong, Z. Chen, M. Fang, L. Zhang, J. Song, D. Yu, Y. Zang, Z. Liu, J. Shi, et al., Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer, Eur. Radiol. 28 (7) (2018) 2772–2778.

[3] H. Su, D. Zhao, H. Elmannai, A.A. Heidari, S. Bourouis, Z. Wu, Z. Cai, W. Gui, M. Chen, Multilevel threshold image segmentation for COVID-19 chest radiography: a framework using horizontal and vertical multiverse optimization, Comput. Biol. Med. 146 (2022) 105618.

[4] A. Qi, D. Zhao, F. Yu, A.A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R.F. Mansour, H. Chen, M. Chen, Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation, Comput. Biol. Med. 148 (2022) 105810.

[5] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, Y. Guo, Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement, Comput. Biol. Med. 147 (2022) 105760.

[6] Z. Wang, Y. Xu, L. Tian, Q. Chi, F. Zhao, R. Xu, G. Jin, Y. Liu, J. Zhen, S. Zhang, A multi-task convolutional neural network for lesion region segmentation and classification of non-small cell lung carcinoma, Diagnostics 12 (8) (2022) 1849.

[7] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A.L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nature Med. 24 (10) (2018) 1559–1567.

[8] T. Nair, J.H. Chuang, et al., The effect of blurring on lung cancer subtype classification accuracy of convolutional neural networks, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2020, pp. 2987–2989.

[9] L. Zhao, X. Xu, R. Hou, W. Zhao, H. Zhong, H. Teng, Y. Han, X. Fu, J. Sun, J. Zhao, Lung cancer subtype classification using histopathological images based on weakly supervised multi-instance learning, Phys. Med. Biol. 66 (23) (2021) 235013.

[10] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, I. Takeuchi, Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3852–3861.

[11] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14318–14328.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[14] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. (2022).

[15] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 2136–2147.

[16] H. Kim, M. Choi, B. Lim, K.M. Lee, Task-aware image downscaling, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 399–414.

[17] F. Saeedan, N. Weber, M. Goesele, S. Roth, Detail-preserving pooling in deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9108–9116.

[18] M. Ehrlich, L.S. Davis, Deep residual learning in the jpeg transform domain, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3484–3493.
[19] L. Gueguen, A. Sergeev, B. Kadlec, R. Liu, J. Yosinski, Faster neural networks straight from jpeg, Adv. Neural Inf. Process. Syst. 31 (2018).
[20] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, F. Ren, Learning in the frequency domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1740–1749.
[21] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 783–792.
[22] Y. Ma, J. Wang, K. Song, Y. Qiang, X. Jiao, J. Zhao, Spatial-Frequency dual-branch attention model for determining KRAS mutation status in colorectal cancer with T2-weighted MRI, Comput. Methods Programs Biomed. 209 (2021) 106311.
[23] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477.
[24] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, Y. Peng, A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data, Irbm 43 (1) (2022) 62–74.
[25] R.J. Chen, M.Y. Lu, J. Wang, D.F. Williamson, S.J. Rodig, N.I. Lindeman, F. Mahmood, Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis, IEEE Trans. Med. Imaging (2020).
[26] N. Braman, J.W. Gordon, E.T. Goossens, C. Willis, M.C. Stumpe, J. Venkataraman, Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 667–677.
[27] Y. Zhang, J. Zhao, Y. Qiang, X. Yang, W. Wu, L. Jia, Improved heterogeneous data fusion and multi-scale feature selection method for lung cancer subtype classification, Concurr. Comput.: Pract. Exper. 34 (1) (2022) e6535.
[28] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, S. Ding, Detecting camouflaged object in frequency domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4504–4513.
[29] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII, Springer, 2020, pp. 86–103.
[30] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-phase shallow learning: rethinking face forgery detection in frequency domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 772–781.
[31] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 14–24.
[32] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 36–46.
[33] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.-C. Chen, Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV, Springer, 2020, pp. 108–126.
[34] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 109–119.
[35] G.K. Wallace, The JPEG still picture compression standard, IEEE Trans. Consum. Electron. 38 (1) (1992) xviii–xxxiv.
[36] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[38] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[40] J. Zhang, X. Zhang, K. Ma, R. Gupta, J. Saltz, M. Vakalopoulou, D. Samaras, Gigapixel whole-slide images classification using locally supervised learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 192–201.
[41] Y. Dong, L. Hou, W. Yang, J. Han, J. Wang, Y. Qiang, J. Zhao, J. Hou, K. Song, Y. Ma, et al., Multi-channel multi-task deep learning for predicting EGFR and KRAS mutations of non-small cell lung cancer on CT images, Quant. Imaging Med. Surg. 11 (6) (2021) 2354.