Explainable Classification of Benign-Malignant Pulmonary Nodules With Neural Networks and Information Bottleneck

Haixing Zhu, Weipeng Liu, *Member, IEEE*, Zhifan Gao, *Member, IEEE*, and Heye Zhang

***Abstract*— Computerized tomography (CT) is a clinically primary technique to differentiate benign-malignant pulmonary nodules for lung cancer diagnosis. Early classification of pulmonary nodules is essential to slow down the degenerative process and reduce mortality. The interactive paradigm assisted by neural networks is considered to be an effective means for early lung cancer screening in large populations. However, some inherent characteristics of pulmonary nodules in high-resolution CT images, e.g., diverse shapes and sparse distribution over the lung fields, have been inducing inaccurate results. On the other hand, most existing methods with neural networks are dissatisfactory from a lack of transparency. In order to overcome these obstacles, a united framework is proposed, including the classification and feature visualization stages, to learn distinctive features and provide visual results. Specifically, a bilateral scheme is employed to synchronously extract and aggregate global-local features in the classification stage, where the global branch is constructed to perceive deep-level features and the local branch is built to focus on the refined details. Furthermore, an encoder is built to generate some features, and a decoder is constructed to simulate decision behavior, followed by the information bottleneck viewpoint to optimize the objective. Extensive experiments are performed to evaluate our framework on two publicly available datasets, namely, 1) the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) and 2) the Lung and Colon Histopathological Image Dataset (LC25000). For instance, our framework achieves 92.98% accuracy and presents additional visualizations on the LIDC. The experiment results show that our framework can obtain outstanding performance and is effective to facilitate explainability. It also demonstrates that this united framework is a serviceable tool and further has the scalability to be introduced into clinical research.**

***Index Terms*— Explainability, information bottleneck, neural network, pulmonary nodule classification.**

Manuscript received 26 September 2022; revised 7 May 2023; accepted 28 July 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1313703, in part by the National Natural Science Foundation of China under Grant 62027813 and Grant 62073118, in part by the Key Research and Development Program of Hebei Province under Grant 21372003D, and in part by the Natural Science Foundation of Hebei Province under Grant F2022202054 and Grant F2020202009. *(Corresponding author: Weipeng Liu.)*

Haixing Zhu and Weipeng Liu are with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment and the School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300130, China (e-mail: zhx930419@163.com; liuweipeng@hebut.edu.cn).

Zhifan Gao and Heye Zhang are with the School of Biomedical Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: gaozhifan@mail.sysu.edu.cn; zhangheye@mail.sysu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3303395.

Digital Object Identifier 10.1109/TNNLS.2023.3303395

I. Introduction

COMPUTERIZED tomography (CT) imaging is a clinically primary and optimally available choice for pulmonary nodules screening and presumptive lung cancer diagnosis. The International Agency for Research on Cancer statistics explicitly manifests that lung cancer remains the highest mortality with 18% of 1.79 million deaths cases and the second highest incidence with 11.4% of 2.2 million new cases worldwide [1]. Lung cancer usually presents as multiple malignant nodules. Research, sponsored by the National Lung Screening Trial, indicates that low dose CT screening can proceed an effective follow-up visit and reduce a 20% mortality [2]. A standard CT screening procedure usually generates hundreds of tomographic images, captured from the mediastinal window and the pulmonary window view.

Manual diagnosis exists such clinical difficulties as time-consuming, variable, and complicated annotation (see Fig. 1). Radiologists read images, adopt a visual observation to scrutinize abnormal regions, record findings that may reveal malignant nodules, perform a diagnostic report and give some suggestions. However, this situation will cause patients to endure a long wait in hospitals with high throughput. In addition, the diagnosis results extremely rely on clinical experience and long-term training [3], and it is also a subjective and inconsistent procedure affected by internal or external factors, i.e., false negatives.

Computer-aided diagnosis, in conjunction with a deep-learning framework, is a desirable solution to complete aided diagnosis. At this stage, deep-learning-oriented pipelines always face the following two challenges (see Fig. 1). *Challenge 1:* Due to the inherent characteristics of pulmonary nodules (e.g., fuzzy boundary, sparse distribution, and subtle differences) in the fine-resolution CT images, it is difficult for models to focus on the discriminable features [4]. *Challenge 2:* The limitations of neural networks are poor explainability and inferior transparency [5]. These alternatives, essentially, are end-to-end black box systems without the necessary components to explain what information is captured, while directly concluding without giving any justifications to support their predictions. Therefore, it is not the best available and trusted solution for radiologists to make a diagnosis. *Motivation:* In light of this, developing an accurate and explainable framework for benign-malignant pulmonary nodule differentiation is a necessary and challenging task.

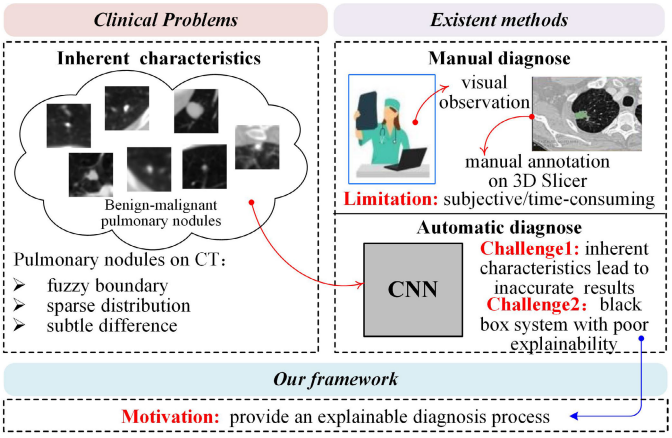

Fig. 1. Overview of existing challenges and research motivation. The challenging difficulties include inherent characteristics (e.g., fuzzy boundary, subtle difference, and spare distribution) of pulmonary nodules in original CT images and the attributes of neural networks (e.g., explainability). In order to overcome these challenges, our motivation aims to build an accurate and explainable framework.

In this article, as is shown in Fig. 1, aiming to improve clinical difficulties and address existing challenges, we propose a unified framework for lung cancer diagnosis, which can differentiate benign-malignant pulmonary nodules and present a visualization to support predictions. Our framework mainly incorporates two stages: 1) image classification and 2) feature visualization (see Fig. 2). In the image classification stage, we construct a two-branch pathway following the refined paradigm [6] to integrate bilateral features. Specifically, inspired by the on–off center-surround (OOCS) [7], we introduce the on-center, off-center, and convolution operation to form a parallel structure, called OOCS-enhanced block, for extending the receptive field. The attention mechanism is applied to focus on the details of local patches. For the feature visualization stage, the information bottleneck, from an information theory viewpoint, is a recent approach to enhance visual explainability and is also a technique to identify relevant features for decisions. Based on the information bottleneck, we also build an encoder–decoder structure to observe what features are utilized, where the encoder aims to generate the feature and the decoder leverages them to perform prediction.

The main contributions are summarized as follows.

1) We develop a unified framework while considering both accuracy and explainability, which facilitates the standard paradigm development in lung cancer diagnosis.
2) We present a novel receptive field module, called OOCS-enhanced block, to strengthen deep-level features learning ability, which aims to extract efficient global features that can contribute to the final prediction.
3) We extend information bottleneck theory to the medical image analysis field, which can provide visual justifications and improve transparency.

A series of experiments are conducted to validate the proposed framework on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI). We obtain competitive performance with extensive quantitative evaluation metrics and qualitatively visual results. To further demonstrate its effectiveness and generalization across different datasets in the lung cancer diagnosis field, we also evaluate our framework on the Lung and Colon Histopathological Image Dataset (LC25000). The promising results on two challenging datasets indicate our framework is effective and viable.

The rest of this article is arranged as follows. Related works about aided-diagnosis pulmonary nodules and explainable approaches are summarized in Section II. Our framework is presented in detail in Section III. The experiment results are reported and analyzed in Section IV. A conclusion is drawn in Section V.

II. Related Works

In this section, we briefly and systematically review typical techniques of pulmonary nodule identification and explainable approaches for medical image analysis.

A. Pulmonary Nodule Classification Methods

Pulmonary nodule benign-malignant classification from CT images has attracted significant concern in recent years. Several systematical reviews are organized to expound on existing challenges and facilitate the state-of-the-art techniques [8], [9]. These approaches are mainly generalized into two aspects: 1) traditional approaches (i.e., nondeep learning) always require professional experience or hand-crafted annotation and 2) deep-leaning-based approaches only rely on labels.

Traditional approaches always involve such means as the threshold [10], region growing [11], morphology [12], statistical models [13], and geometrical-level sets [14]. These frameworks usually adopt specific techniques, including threshold, region growing, and level sets, to generate candidate regions, where the features are extracted (i.e., texture, shape, and size) by manual or semiautomatic means, followed by feeding into a specific classifier to perform classification or regression tasks. However, this procedure is not powerful and efficient to address the subtle shape variation, especially the nodules that have high similarity with blood capillary, and it may introduce extrinsic factors that have a negative effect on classification results. A prerequisite is prior knowledge and domain experience.

Recently, deep-learning-based approaches have been explored to develop a high-efficiency and accuracy architecture in the pulmonary nodule analysis field. Shen et al. [15] developed a unique pooling operation and directly model raw nodule patches by the constructed multicrop convolutional neural network (CNN) for lung nodule malignancy suspiciousness evaluation. Shabi et al. [16] investigated a progressive learning algorithm by introducing a Bernoulli matrix and gradually adding easy and difficult samples into the training process for nodules classification. In [17], an expert system is constructed by gathering multiparallel capsule networks, and a gating mechanism is introduced to pay attention to correlations of different subnetworks for predicting malignant nodules.

The 3-D region proposal network is a popular measure to deal with tasks in malignancy probability estimation [18] or false positives reduction [19]. Related works have employed

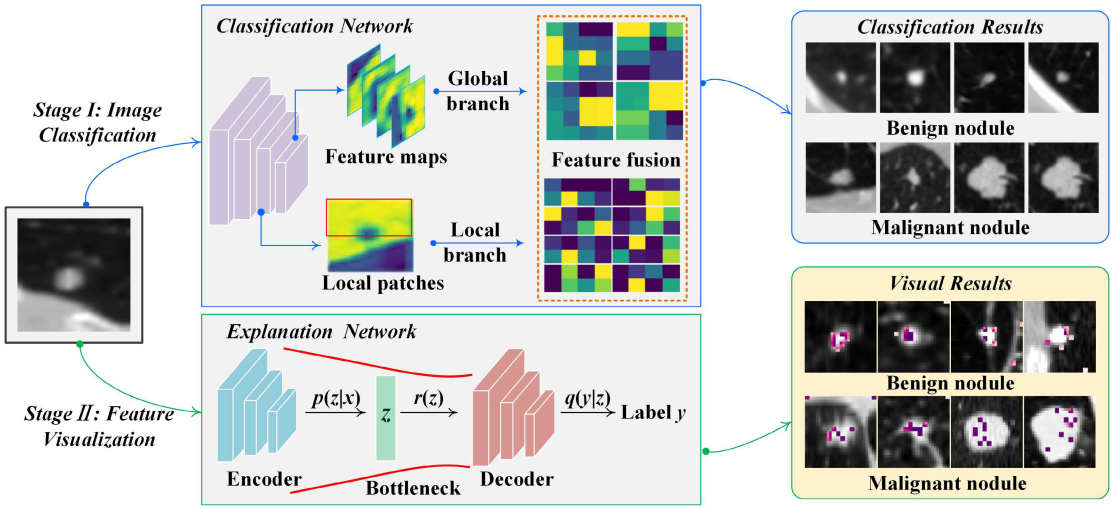

Fig. 2. Whole workflow of our united framework. It mainly incorporates two stages: image classification and feature visualization. The former is driven by a global-local refinement structure, and the latter constructs an encoder–decoder to parameterize the variational information bottleneck for providing visual explanations.

multifarious 3-D components to learn multidimensional representation from 3-D views [20], [21], [22], [23], [24], 2-D views [25], [26], and a group of patches [27]. Xie et al. [20] designed nine knowledge-based collaborative subnetworks to extract multiview 3-D features and incorporate them for classifying nodules. Setio et al. [22] constructed a multistream CNN to learn features from CT patches and fuse them using mixed feature fusion strategies. Some latest techniques including generative adversarial network (GAN) [28], multiple kernel learning [29], data-driven [30], adversarial models [31], multitasks learning [32], and searching and pruning [33] also are applied to solve the limitation in data shortage and class imbalance.

Controlling the receptive field is an important and familiar measure to strengthen the learning ability of networks. It imitates the human visual perceptual system to traverse whole regions and generate a batch of feature maps, followed by integrating information to perform representation learning [34], medical image classification [35], and detection [36], [37], [38], [39] and segmentation tasks [40], [41], [42], [43], [44]. Some works adopt various dilated convolution [35], [36], deconvolution [37], and project well-designed blocks [38], [39] to learn deep-level or multiscale features. To avoid omitting spatial information, [35] and [40] employed dilated convolutions with different dilation rates to construct a pyramid construction. In addition, Dou et al. [38] and Zhao et al. [39] set a multibranch network with different receptive fields to cover as many features as possible while reducing as much redundant information as possible. Recently, [42], [43], and [44] indicated that transformer-based frames, benefiting from a self-attention mechanism, can generate large-scale receptive fields to capture effective information from long-distance or large-range images. However, these approaches may contain much unnecessary information. To alleviate this limitation, a novel receptive field named OOCS is proposed in [7] and has been successfully applied to medical image 3-D segmentation [45]. The above approaches only pay attention to pursuing precision and ignore explainability.

B. Explainability for Medical Images Analysis

Explainability is a crucial criterion in building, improving, and deploying neural networks in medical fields. There is yet a unified and standard frame to seek a balance between accuracy and explainability. Most works apply heatmap [46] and its multiple variants [47], [48] to provide visual explainability and transparency in the postprocessing stage [49], [50], [51], [52], [53]. Driven by the attention-guided CNN, Guan et al. [49] trimmed a discriminative region by generating corresponding heatmaps to imitate the specialistic CT scan reading procedure. Zhang et al. [50] employed class activation mapping (CAM) to visualize the lesion regions on which attention modules focus. Furthermore, Oh et al. [51] introduced the probability value of the corresponding CAM. Tang et al. [52] combined guided Grad-CAM with confidence maps to locate and highlight specific regions. However, these methods can produce coarse-grained explanations.

Two representative works in [54] and [55] manifest that promoting explainability by joint diagnosis reports and medical images is an optional solution. Zhang et al. [54] constructed three cooperative subnetworks, a scanner for detecting lesions from pathological images, a diagnoser for drawing the outline of the regions-of-interest and generating explicable reports, and an aggregator to integrate information and perform diagnosis results. In [55], the ResNet network is utilized to detect cancer, and a long–short-term memory (LSTM) network is implemented to enhance visual relatedness from reports to images. This kind of approach is complicated because they require joint diagnosis reports and medical images.

There are also some related techniques to evaluate explainability. Tschandl et al. [56] promoted an interactive interface powered by an image retrieval technique, and doctors can search similar images and associated reports from databases to support an explainable diagnosis. Gu et al. [5] designed an

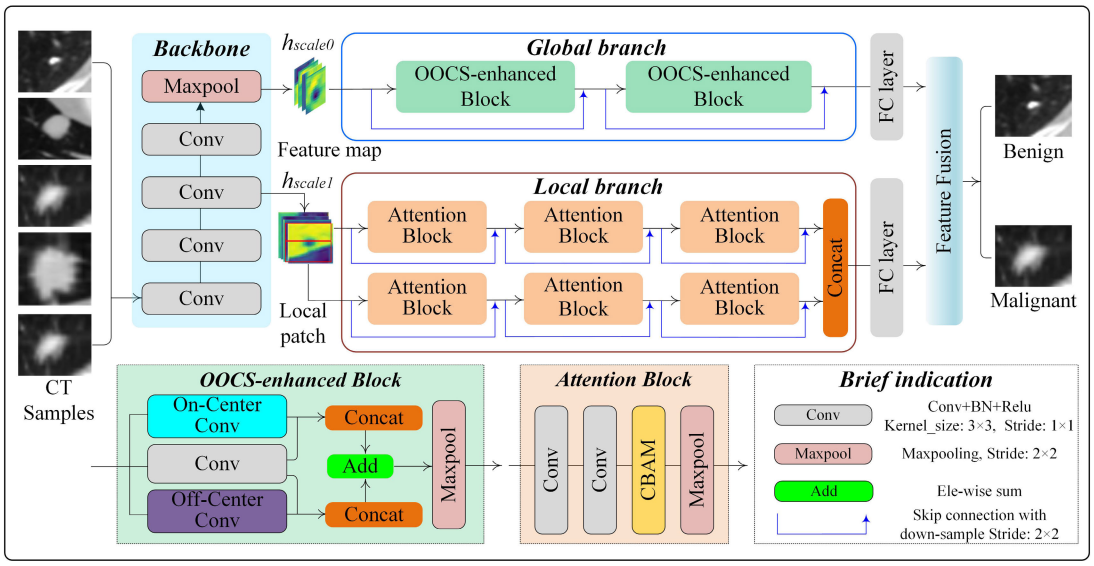

Fig. 3. Entire structure of the classification network. It includes three parts: a backbone is promoted to generate multilevel features from CT images, a global branch is composed of two OOCS-enhanced blocks to extract deep-level features from larger feature maps, and a local branch is built by attention blocks to focus on the detailed information from patches of the smaller feature maps.

importance estimation network to generate a visual explanation for eliminating irrelevant features and providing classification reference. Shen et al. [57] composed an explainable hierarchical semantic CNN (HSCNN) to output two levels of results: deep semantic information for classification and a piece of shallow information for presenting explainability.

Information bottleneck, an information theory viewpoint, has been applied to explore the model's explainability, especially in natural language processing [58], [59]. Recently, it is also advanced to present visual effects in medical fields [60], [61]. Wang et al. [60] generated segmentation masks and estimated which regions are activated for breast cancer diagnosis. Khakzar et al. [61] combined feature attribution with the information bottleneck to estimate the importance of input for making decisions.

Discussion for Related Works: It is possible and necessary to seek a tradeoff between accuracy and interpretability. Higher accuracy requires the classification network to learn representative and efficient features for decisions. Better interpretability means the classification models need to provide visual or reliable evidence for supporting the corresponding decision. Although some references have been performed to promote explainable classification, these techniques do not considered both accuracy and interpretability. In our work, we proposed a united framework to enhance accuracy and improve transparency.

III. Methods

A. Overview

The whole workflow includes two components, i.e., classification network (see Section III-B) and explanation network (see Section III-C), as is shown in Fig. 2. Specifically, we first feed the processed CT images with 32 × 32 resolution into the classification network to generate multilevel feature maps. Then, the classification process is powered by establishing a two-branch pathway to learn global-local features from multilevel feature maps, respectively, as is shown in Fig. 3. Furthermore, instead of only performing a benign-malignant binary classification task, our framework also includes an explanation network driven by the information bottleneck approach, so that to provide informative and visual information.

B. Classification Network

Considering the fuzzy boundary, sparse distribution, and subtle differences of pulmonary nodules in the original fine-resolution CT images, we engineer global and local branches to unite discriminative features for benign or malignant nodule diagnosis. Let $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{(H,W)}$ denotes the processed images, where H and W are the height and width of input images. $y = (y_b, y_m)$ denotes the class label, where y_b and y_m with value 0 or 1 represents benign and malignant nodules.

Fig. 3 shows the well-designed blocks and integrated structure in detail. Specifically, this stage takes processed CT images with 32 × 32 resolution as input and starts with four specific convolution and maxpooling operations as the backbone to generate coarse and hierarchical feature maps with different resolutions as the base features, e.g., h_{scale0} and h_{scale1}, which are designed to provide multilevel information for follow-up two-branch stage, respectively. Each built-up layer in the backbone architecture contains one convolution layer with followed batch normalization layer and rectified linear unit (ReLU) activation function. The first four convolutional layers are composed of 3 × 3 kernel filters with stride 1 and padding 2, and the fourth convolutional is followed by a maxpooling layer with stride 2.

Inspired by the visually bio-mimetic motif, according to the OOCS pathways mentioned in Section II-A [7], an extension called OOCS-enhanced block is constructed to organize a global branch, which can enlarge the receptive field and

enhance the extraction ability of edge information. More specifically, as is shown in Fig. 3, we introduce on-center, off-center, and conventional convolution to formulate a parallel scheme that can expand the receptive field without downsampling, followed by the element-wise and concatenation operation integrate multipath features step-by-step. Furthermore, we take two cascaded OOCS-enhanced blocks to perform coarse-to-fine extraction. Skip connections also are adopted to adequately integrate multilevel features and enrich information flow inside the network. The feature map h_{scale0} is taken as input into the global branch. Let g denotes the global feature learning, the h_{scale0} is taken as input and extracted deeply global feature h_{global}

$$h_{\text{global}} = g(h_{\text{scale0}}). \tag{1}$$

In the local branch, we divide the feature maps h_{scale1} into two patches along the horizontal axis. As shown in Fig. 3, each attention block comprises two convolutional layers, one convolutional block attention module (CBAM), and one maxpooling layer. The local branch is also driven by a parallel pathway, where each pathway includes three cascaded attention blocks. To further enrich semantic information and mitigate parameters, we add skip connections in the downsampling path and utilize element-wise operations to fuse features. Let f denotes the local feature learning, and we feed two patches into the local branch to extract the detailed features h_{local}

$$h_{\text{local}} = f(h_{\text{scale1}}). \tag{2}$$

In order to conduct global-local features aggregation, we adopt a decision-level fusion strategy to aggregate global-local features [62]. The fully connected layer is followed by a softmax layer to formulate h_{global} and h_{local} into the probability of benign or malignant pulmonary nodules. The cross-entropy is chosen to calculate classification loss. Our loss function contains two parts: $\mathcal{L}_{\text{global}}$ and local branch $\mathcal{L}_{\text{local}}$, which can be expressed as

$$\begin{cases} \mathcal{L}_{\text{global}} = -\left(y \log p^1_{\text{global}} + (1-y) \log p^0_{\text{global}}\right) \\ \mathcal{L}_{\text{local}} = -\left(y \log p^1_{\text{local}} + (1-y) \log p^0_{\text{local}}\right) \end{cases} \tag{3}$$

where p_{global} and p_{local} are the probability value of the softmax layer from different branches; p^0 and p^1 indicate labels with the value 0 and 1 corresponding to being or malignant nodules, respectively.

The final classification loss function is defined as follows:

$$\mathcal{L} = \eta \mathcal{L}_{\text{global}} + (1-\eta)\mathcal{L}_{\text{local}} \tag{4}$$

where η denotes a balance factor of the two cross-entropy losses.

C. Explanation Network

1) Variational Information Bottleneck: In this section, we apply the variational information bottleneck, an information theory viewpoint, to enhance explainability by providing visual results (see Fig. 2). The information bottleneck-based approach aims to seek an optimal representation z by maximally compressing information as much as possible from input instance x to representation z, and maximally preserve information as much as possible from representation z to label y [63]. The information bottleneck objective is to optimize the following equation:

$$\mathcal{L} = I(z, y) - \beta I(z, x) \tag{5}$$

where $I(z, y)$ and $I(z, x)$ denote the mutual information. β denotes the tradeoff value to adjust the mutual information between z and y.

Since it is intractable to directly optimize the objective loss in (5), variational mutual information is leveraged to derive new bounds in [64]. The variational bound $I(z, y)$ can be expressed in the following equation:

$$I(z, y) = \int p(y, z) \log \frac{p(y|z)}{p(y)} dydz \tag{6}$$

where $p(y|z)$ is difficult to calculate $I(z, y)$. The $q(y|z)$ is devised to approximately estimate $p(y|z)$. Furthermore, the Kullback–Leibler divergence is used to approximate $p(y|z)$

$$\text{KL}\big[p(y|z), q(y|z)\big] \geq 0 \tag{7}$$

hence,

$$\int p(y|z) \log p(y|z) dy \geq \int p(y|z) \log q(y|z) dy \tag{8}$$

and leads to

$$I(Z, Y) \geq \int p(y, z) \log q(y|z) dydz + H(Y) \tag{9}$$

where $H(Y)$ is the entropy of label y and can be ignored.

Then, the variational bound $I(z, x)$ is defined as follows:

$$I(z, x) \leq \int p(x) p(z|x) \log \frac{p(z|x)}{r(z)} dxdz. \tag{10}$$

2) Variational Information Bottleneck for Explainability: As shown in Fig. 2, an encoder–decoder structure is powered to parameterize $p(z|x)$ and $q(y|z)$. The encoder takes $X = (x_1, x_2, \ldots, x_n)$ as input to generate a compressed representation z. Then, we utilize a softmax layer to transform them into corresponding class probabilities $p(z|x)$. Following the process in language understanding task [58], we define $z = z \odot X$. After that, we feed z into the decoder which is regarded as a classifier to perform the prediction task and output y.

Here, the bound $I(z, y)$ can be derived according to (9) as follows:

$$I(z, y) \geq \mathbb{E}_{z-p(z|x)} \log q\big[(y|z \odot X)\big]. \tag{11}$$

Meanwhile, according to (10), the bound $I(z, x)$ can be expressed

$$I(z, x) = \beta \text{KL}\big[p(z \odot X|x), r(z \odot X)\big]. \tag{12}$$

Combining with two variational bounds (11) and (12), the objective is established as follows:

$$\mathcal{L}_{\text{VIB}} = \mathbb{E}_{z-p(z|x)}\big[\log q(y|z \odot X)\big] + \beta \text{KL}\big[p(z \odot X|x), r(z \odot X)\big]. \tag{13}$$

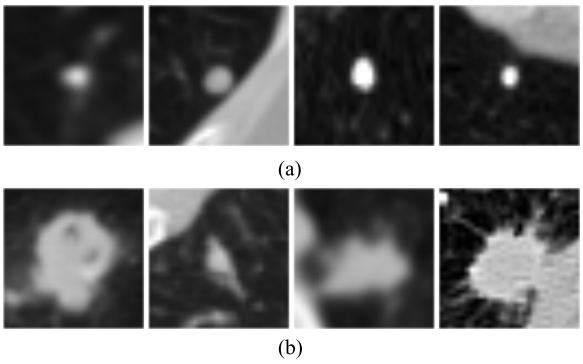

Fig. 4. Sample of preprocessed pulmonary nodule images on the LIDC dataset. (a) Benign. (b) Malignant.

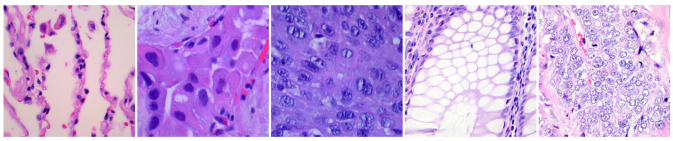

Fig. 5. Samples of preprocessed LC25000 dataset. (a) Benign lung tissue (Lungn). (b) Lung adenocarcinomas (Lungac). (c) Lung squamous cell carcinomas (Lungscc). (d) Benign colon tissue (Colonn). (e) Colon adenocarcinomas (Colonac).

IV. Experiments

A. Datasets

We evaluate our framework on two public datasets with different imaging modalities and characteristics for cancer diagnosis: 1) CT-based classification task estimates benign or malignant pulmonary nodules on the LIDC dataset and 2) pathology images-based classification task differentiates benign or malignant tissue.

1) Lung Image Database Consortium and Image Database Resource Initiative: The LIDC-IDRI dataset is sponsored by the National Cancer Institute for monitoring early lung cancer in high-risk groups [65]. Researchers collect 1018 CT instances that include 1010 patients' thoracic CT scans and correspondingly annotated lesions. Each annotation is performed by four experienced thoracic radiologists through a two-phase annotation process. For each nodule, radiologists independently observe images and divide them into three classes according to diameters: nodules $>$ or $=$ 3 mm, nodules $<$ 3 mm, and no nodules. We choose nodule diameters $>$ or $=$ 3 mm as alternative samples and remove diameters nodules diameters $<$ 3 mm and no nodules. Fig. 4 shows some pretreated samples of benign-malignant pulmonary nodules in the LIDC.

In addition, radiologists describe each nodule with multiple characteristics such as subtlety and malignancy in the annotation. All malignant nodules are sorted into ranges from 1 to 5, in which 1 represents the benign nodules and 5 represents the malignant nodules. In our work, according to the specified principle in [20], we group nodules into two categories: for nodules that a malignancy greater than 3, we divide them into malignant nodules and set labels as 1; for nodules that a malignancy less than 3, we define them into benign nodules and set label as 0; for nodules that a malignancy equal to 3, we state them as uncertain nodules and not take them as alternative samples. To sum up, it contains 442 benign nodules and 406 malignant nodules in our experiments.

2) Lung and Colon Histopathological Image Dataset (LC25000): LC25000 is collected to develop advanced techniques for pathological image analysis in [66]. This dataset includes five classes of histopathological images, such as lung benign tissue (Lungn), lung adenocarcinoma (Lungac), lung squamous cell carcinoma (Lungscc), colon adenocarcinoma (Colonac), and colon benign tissue (Colonn), as shown in Fig. 5. Each class has 5000 samples with 768 $\times$ 768 pixels. We employ the entire dataset for the following experiments, a total of 25 000 histopathological images.

B. Preprocessing

The raw chest CT images involve the whole lung, pleura, and trachea, while nodules always present small, subtle, and sparsely distributed over the whole lung fields. The scanning procedure manufactures CT images with high resolutions and contains a large number of useless regions. Therefore, it is necessary to perform preprocessing to remove needless portions. First, we segment the lung parenchyma and clip the irrelevant information. Second, we utilize annotation to read each nodule's information and divide them into two classes: benign and malignant. Finally, 846 benign-malignant nodule images with 32 $\times$ 32 resolutions are made as training and test data. For the LC25000 dataset, we only resize the original images as 64 $\times$ 64 resolutions.

C. Implementation Details and Evaluation Metric

In our experiments, two datasets are randomly split into the training sets, validation sets, and test sets with 60%, 20%, and 20% of total images, respectively. The proposed framework is implemented on the PyTorch environment with NVIDIA GeForce 2080Ti GPUs. For the image classification process, we employ the stochastic gradient descent (SGD) optimizer to train the network, starting with an original learning rate of 0.001 and a decay coefficient of 1×10^{-4}. We set the batch size as 32 and the number of train epochs as 50. For the following explanation network, the adaptive moment estimation (Adam) is adopted with the learning rate of 0.0001 and the other default parameters. We also set the batch size as 32, while the number of epochs is set as 100.

We evaluate our framework under various metrics in two different stages as follows.

1) *Image Classification Stage:* We calculate six quantitative metrics to evaluate the classification performance, which includes the Accuracy $=$ (TP $+$ TN)/(TP $+$ TN $+$ FP $+$ FN), Precision $=$ TP/(TP$+$FP), Sensitivity $=$ TP/(TP$+$FN), Specificity $=$ TN/(TN $+$ FP), F1-score $=$ (2 $\times$ Precision$\times$Sensitivity)/(Precision$+$Sensitivity), and the area of value under the receiver operating characteristic curve (AUC), where TN, TP, FN, and FP denote the value of true negatives, true positives, false negatives, and false positives, respectively.

TABLE I
ABLATION ANALYSIS OF DIFFERENT COMPONENTS IN CLASSIFICATION NETWORK ON THE LIDC AND LC25000 DATASETS BOLD INDICATES THE BEST PERFORMANCE

Dataset	Methods		Accuracy(%)	AUC(%)	Sensitivity(%)	Specificity(%)
	Global branch	Local branch				
LIDC	✓		90.39	96.37	87.68	92.87
		✓	91.86	97.35	89.16	94.34
	✓	✓	**92.98**	**97.55**	**91.26**	**94.57**
LC25000	✓		96.78	99.77	96.77	99.19
		✓	94.1	99.47	94.09	98.52
	✓	✓	**97.52**	**99.9**	**97.53**	**99.38**

TABLE II
ABLATION ANALYSIS OF PARAMETER SETTING AND FEATURE FUSION METHOD IN CLASSIFICATION NETWORK ON THE LIDC AND LC25000 BOLD INDICATES THE BEST PERFORMANCE

Dataset	Parameter η	Accuracy(%)	Precision(%)	Sensitivity(%)	Specificity(%)	AUC(%)	F1-Score(%)
LIDC	0.3	92.92	93.14	92.0	93.78	**97.79**	92.57
	0.5	**92.98**	**93.92**	91.26	**94.57**	97.55	**92.57**
	0.7	92.28	90.78	**93.35**	91.29	97.54	92.05
	Feature-level	89.86	88.1	91.13	88.69	96.44	89.59
LC25000	0.3	97.28	97.28	97.29	99.32	99.83	97.27
	0.5	**97.52**	**97.52**	**97.53**	**99.38**	**99.9**	**97.51**
	0.7	97.44	97.47	97.44	99.36	99.85	97.45
	Feature-level	95.66	95.65	95.66	98.91	99.71	95.64

2) *Feature Visualization Stage:* We also validate the explanation network with the Accuracy, Precision, Recall, and F1-score.

D. Experiments

In order to verify the efficacy and evaluate the performance of our framework, we carry out a series of experiments to enumerate quantitative and qualitative results. The experiments are mainly performed in two aspects as follows. 1) Ablation analysis: we conduct ablation experiments to validate different components and strategies. 2) Comparison experiments: we also compare our classification network against advanced approaches.

First, to validate the contribution of different components, we conduct an ablation experiment on the LIDC and LC25000 datasets. Specifically, the backbone is taken as a baseline, and append the rest components to it step by step so that we can divide them into three control groups: backbone + global branch, backbone + local branch, and backbone + global branch + local branch. By setting the comparison with distinct combining forms, quantitative values of the accuracy, sensitivity, specificity, and AUC are reported in Table I for ablation analysis.

Second, to scrutinize the benefits of two fusion strategies, we organize comparison experiments to understand the advantage of the decision-level approach. Distinguished from immediate fusion, decision-based is necessary to introduce an important hyper-parameter η in the loss function. We then set control groups to investigate the influence of η by setting values as 0.3, 0.5, and 0.7. Comprehensive evaluations are reported via opting accuracy, precision, sensitivity, specificity, F1-score, and AUC in Table II.

Third, aiming to evaluate the overall performance of our classification framework, we perform the following comparison experiments on two datasets and report results in Tables III and IV. All comparative approaches can be grouped into three aspects as follows.

1) *Classical Image Classification Approaches:* These approaches first are advanced to handle nature image analysis tasks and also appear in the medical field. We select the two representative models (e.g., VGG16 and ResNet18) and compare them against our framework on the LIDC and LC25000 datasets.
2) *Literature Approaches for Pulmonary Nodule Classification:* These approaches employ several learning frameworks to extract features that present heterogeneity, including texture and shape features [67], [68]. Some works design specialized blocks according to the characteristics of CT images, such as multicrop strategy [15], MIXCAPS [17], HSCNN [57], deep residual network [69], global-local network [70], and 3-D CNN [24], to measure multiview or multilevel features for refining performance.
3) *Machine-Learning-Based Approaches for Histopathological Image Classification:* To further demonstrate the generalization ability in the related tasks, we extend

TABLE III
COMPARISON OF DIFFERENT APPROACHES FOR BENIGN-MALIGNANT PULMONARY NODULE CLASSIFICATION ON THE LIDC BOLD INDICATES THE BEST PERFORMANCE

Methods	LIDC Dataset					
	Accuracy(%)	Precision(%)	Sensitivity(%)	Specificity(%)	AUC(%)	F1-Score(%)
VGG16	86.79	93.41	78.57	94.91	96.4	85.35
ResNet18	88.56	83.59	**94.7**	82.92	96.5	88.8
Texture+shape features [67]	-	-	89.73	86.36	-	-
Deep visual features [68]	86.79	-	60.26	**95.42**	-	--
Multi-crop CNN [15]	87.14	-	77.0	93.0	93.0	-
MIXCAPS [17]	90.7	-	89.5	93.4	95.6	-
HSCNN [57]	84.2	-	70.5	88.9	85.6	-
Deep residual network [69]	89.9	-	91.07	88.64	94.59	-
Global-local network [70]	88.46	87.38	88.66	-	95.62	-
3D CNN [24]	90.6	-	83.7	93.9	93.9	-
Ours	**92.98**	**93.92**	91.26	94.57	**97.55**	**92.57**

TABLE IV
COMPARISON OF DIFFERENT APPROACHES FOR HISTOPATHOLOGICAL IMAGES CLASSIFICATION ON THE LC25000 BOLD INDICATES THE BEST PERFORMANCE

Methods	LC25000 Dataset
	Accuracy(%)
Statistical features + RF [71]	87.2
GLCM + SVM [71]	90.0
Hu invariant moments + LDA [71]	62.0
Statistical features + GLCM + RF [71]	94.2
GLCM + Hu invariant moments + SVM [71]	89.0
VGG16	95.38
ResNet18	95.26
Ours	**97.52**

TABLE V
ANALYSIS OF EXPLANATION NETWORK WITH DIFFERENT PARAMETER β ON THE LIDC AND LC25000 BOLD INDICATES THE BEST PERFORMANCE

parameter	β	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
LIDC	0.01	84.85	**90.88**	75.57	81.89
	1	**85.14**	90.62	**76.27**	**82.15**
	100	84.79	90.05	76.12	81.88
LC25000	0.01	81.32	79.77	79.69	77.32
	1	**81.88**	**80.64**	**81.15**	**78.52**
	100	81.22	79.69	80.26	77.5

comparison experiments on the LC25000 dataset and compare it with results incorporated from the relevant published papers [71]. These techniques are acquired by means of features engineering, e.g., statistical features, gray level co-occurrence matrix (GLCM) and Hu invariant moments, and machine-learning classifiers, e.g., support vector machine (SVM), random forest (RF), and linear discriminant analysis (LDA). It should be noted that the LC25000 has balanced class attributes and multimetric evaluation results are approximate, therefore we only utilize the accuracy to quantificationally measure performance.

Finally, to show the availability and feasibility of the feature visualization, we verify the ability of the encoder–decoder structure, which is employed to parameterize the variable information bottleneck. Concretely, the metrics of accuracy, precision, recall, and F1-score are reported to achieve quantitative evaluation in Table V. Besides, we also present the qualitative analysis to explain which regions are focused for pulmonary nodules or pathology image identification.

E. Results and Discussion

1) Effectiveness of the Different Components: The pipeline driven by a global-local branch shows superior performance than a single-branch framework. As is shown in Table I, the experiment results (see third and sixth rows) on the LIDC and LC25000 indicate the bilateral structure can effectively accelerate an improvement compared to the single-branch pipeline. For instance, our classification network achieves quite promising performance on the LIDC (accuracy: 92.98%, AUC: 97.55%, sensitivity: 91.26%, and specificity: 94.57%) and outperforms the single global-branch (2.87%, 1.22%, 4.08%, and 1.83%), the single local-branch (1.22%, 0.21%, 2.36%, and 0.24%). This phenomenon manifests that these components can integrate multilevel feature information to enhance network learning ability, and also elucidates that global-local refinement structure is better complementary to some extent.

2) Analysis of the Feature Fusion Strategy: The decision fusion-based measure can make a better tradeoff between global and local information. According to quantitative results reported in Table II, the comparative points present decision-level fusion is superior to feature-level fusion, for

example, the outcomes on the LIDC dataset can achieve better performance with the accuracy of 92.98%, precision of 93.92%, sensitivity of 91.26%, specificity of 94.57%, AUC of 97.55%, and F1-score of 92.57% when η is set as 0.5, and outweighs the feature-level fusion by 3.47%, 6.61%, 0.14%, 6.63%, 1.15%, and 3.33%. In addition, we also explore the influence of different parameter values η, which is introduced for balancing the global-local feature information. We select the range of η from 0.3 to 0.7 and give the results of three values (η is set as 0.3, 0.5, and 0.7). As is shown in Table II, we find that the influence of parameter η is slight. It can achieve a satisfying performance when the η is set as 0.5.

3) Comparison to State-of-the-Art Approaches: Compared with the various approaches mentioned in Section IV-D, our classification framework achieves leading and comprehensive strengths. Table III lists the numerical results.

1) Comparison to classical approaches: we can observe in Table III that ResNet18 achieves the optimal sensitivity with the value of 94.7%. However, the accuracy (VGG16: 86.79% and ResNet: 88.56%) is not satisfactory.
2) Comparison to literature approaches: it can be noticed that our framework achieves eminent grades using almost all evaluation metrics. Specifically, the texture and shape-based approach merely obtains the non-competitive values of sensitivity (89.73%) and specificity (86.36%). It is worth mentioning that the deep visual features-based approach has the highest specificity of 95.42% under the help of priori knowledge. The multicrop method does not perform well only with the accuracy of 87.14%, even lowering the standard of single-branch structure (see the second row from Table I). The ensemble MIXCAPS approach presents optimal performance except for our framework. In comparison with the results of MIXCAPS, the accuracy, AUC, sensitivity, and specificity increased by approximately 2.51%, 2.04%, 1.97%, and 1.25%, respectively. Since the authors of these methods have not released their codes, we thus compare our results with those reported in their papers.
3) Comparison to machine-learning-based approaches: this kind of approach requires combining feature engineering with classifiers. We can observe in Table IV that these approaches have poor performance on the LC25000 dataset. For example, the framework involving statistical features and RF only gets an accuracy of 87.2%, even much less than the value achieved by our framework (97.52%). The reason may be that it is inevitable to import interferential noises and redundant information during feature descriptions. It is fallible and sophisticated to calculate features and feed features into classifiers, while our framework addresses the above limitations by achieving a fully automatic feature extraction.

Tables III and IV show that the proposed framework realizes significant performance on the relevant classification tasks, which has a preferable capability compared with the other state-of-the-art approaches. The enrichment is mainly ascribed to the global-local retained structure, which enables a deep neural network to pay attention to both global high-level and local detailed features and hence strengthens the network's ability to extract prominent information. In addition, Fig. 6 shows the confusion matrices of the three approaches on two datasets. The visual results also indicate our framework obtains excellent performance.

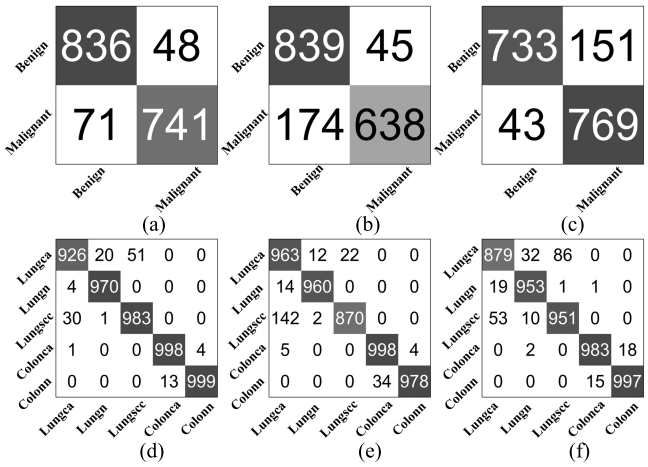

Fig. 6. Confusion matrices of the different approaches on the LIDC and LC25000 datasets. (a) and (d) Ours. (b) and (e) VGG16. (c) and (f) ResNet.

4) Analysis of the Explanation Network: The feature explainable framework based on the variational information bottleneck is driven and parameterized by an encoder–decoder structure. We evaluate the encoder–decoder under the metrics of accuracy, precision, recall, and F1-score. Table V shows the performance on the LIDC datasets with the accuracy of 85.14%, precision of 90.62%, recall of 76.27%, and F1-score of 82.15% when β is set as 1. It reflects that our framework is effective to imitate the classification performance. We also pick up experiment data on the LC25000 dataset with an accuracy of 81.88%, precision of 80.64%, recall of 81.15%, and F1-score of 78.52%. It is noticeable that the predictions ($\beta = 1$) outperform the other results (when $\beta = 0.01$ and 100).

5) Visualization: Fig. 7 shows the visual results generated by the explanation network on two challenging datasets. The highlighted pixels prove that the features are activated and utilized as effective information to make classification decisions. As is shown in Fig. 7(a), we observe that the classification network can pay attention to the nodules region in processed CT images. In Fig. 7(b), we analyze that the classification network captures informative pixels, in particular those containing distinct features that can distinguish various lung and colon diseases.

F. Limitations

This work has potential limitations. The high-resolution CT images exist in some irrelevant but similar tissues, i.e., pulmonary blood capillary, especially those in an adjacent or contiguous relationship with pulmonary nodules. Our framework focuses on 2-D features and ignores spatial information, and this situation may extract interferential features and cause

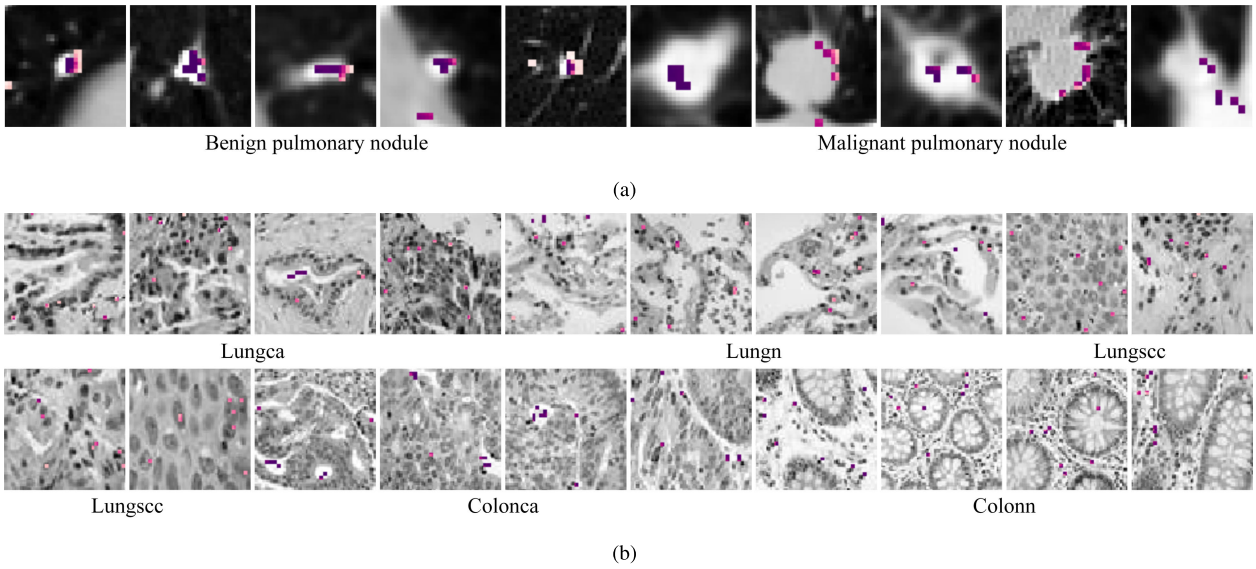

Fig. 7. Some samples of visualization explanation results of benign-malignant pulmonary nodules on the LIDC dataset and pathology image classification on the LC25000. (a) LIDC. (b) LC25000.

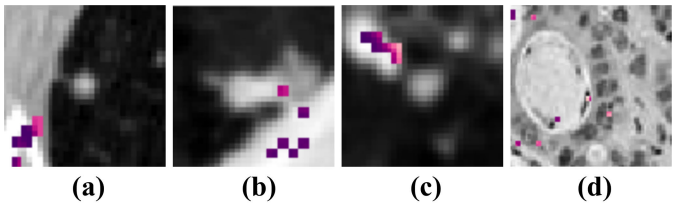

Fig. 8. Some failure samples. (a) Benign nodules while (b) and (c) are malignant nodules, and (d) is lung adenocarcinoma, but our framework focuses on the irrelevant tissues for classification.

incorrect results. Besides, though our explanation network can present observable and visual markers in regular images (see Fig. 7), this process also generates frustrating results in challenging or difficult images (see Fig. 8). This phenomenon indicates that our framework extracts disrelated information. In future work, we will make full use of 3-D and continuous information on CT sequences.

V. Conclusion

In this work, we propose a unified framework with a classification network and an explanation network to differentiate benign-malignant pulmonary nodules and improve explainability. The classification network is powered by a dual branch-oriented paradigm, where the global branch extracts deep-level features from the preprocessed images and the local branch follows the detailed information from patches of CT images. Meanwhile, we jointly build an encoder–decoder structure to parameterize information bottleneck in the feature visualization stage, so that provides visual judgments for diagnosis. We perform a series of experiments on two challenging datasets about lung cancer diagnosis, i.e., LIDC and LC25000, to demonstrate that our framework is valid and feasible. The competitive performance shows it has the potential capacity to provide an auxiliary diagnosis, and has the scalability to alleviate clinical pressure with high throughput cases. Therefore, our future work will delve into more CT image information to improve accuracy and extend additional components to augment multitask analysis, e.g., quantitative calculation of nodules size.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A Torre, and F. Jemal, "GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] D. Aberle et al., "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England J. Med.*, vol. 365, no. 5, pp. 395–409, 2011.

[3] W. M. Reed, J. T. Ryan, M. F. McEntee, M. G. Evanoff, and P. C. Brennan, "The effect of abnormality-prevalence expectation on expert observer performance and visual search," *Radiology*, vol. 258, no. 3, pp. 938–943, Mar. 2011.

[4] N. Khosravan, H. Celik, B. Turkbey, E. C. Jones, B. Wood, and U. Bagci, "A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning," *Med. Image Anal.*, vol. 51, pp. 101–115, Jan. 2019.

[5] D. Gu et al., "VINet: A visually interpretable image diagnosis network," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1720–1729, Jul. 2020.

[6] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101908.

[7] Z. Babaiee, R. Hasani, M. Lechner, D. Rus, and R. Grosu, "On-off center-surround receptive fields for accurate and robust image classification," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 478–489.

[8] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis—A survey," *Pattern Recognit.*, vol. 83, pp. 134–149, Nov. 2018.

[9] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[10] C. Jacobs et al., "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Med. Image Anal.*, vol. 18, no. 2, pp. 374–384, Feb. 2014.

[11] J. Song et al., "Lung lesion extraction using a toboggan based growing automatic segmentation approach," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 337–353, Jan. 2016.

[12] W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical ct images," *IEEE Trans. Med. Imag.*, vol. 22, no. 10, pp. 1259–1274, Oct. 2003.

[13] A. A. Farag, H. E. A. E. Munim, J. H. Graham, and A. A. Farag, "A novel approach for lung nodules segmentation in chest CT using level sets," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5202–5213, Dec. 2013.

[14] K. Guo, X. Liu, N. Q. Soomro, and Y. Liu, "A novel 2D ground-glass opacity detection method through local-to-global multilevel thresholding for segmentation and minimum Bayes risk learning for classification," *J. Med. Imag. Health Informat.*, vol. 6, no. 5, pp. 1193–1201, Sep. 2016.

[15] W. Shen et al., "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognit.*, vol. 61, pp. 663–673, Jan. 2017.

[16] M. Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108309.

[17] P. Afshar, F. Naderkhani, A. Oikonomou, M. J. Rafiee, A. Mohammadi, and K. N. Plataniotis, "MIXCAPS: A capsule network-based mixture of experts for lung nodule malignancy prediction," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107942.

[18] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-OR network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3484–3495, Nov. 2019.

[19] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, "SANet: A slice-aware network for pulmonary nodule detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4374–4387, Aug. 2022.

[20] Y. Xie et al., "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 991–1004, Apr. 2019.

[21] A. A. A. Setio et al., "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[22] A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.

[23] P. Zhai, H. Cong, E. Zhu, G. Zhao, Y. Yu, and J. Li, "MVCNet: Multiview contrastive network for unsupervised representation learning for 3-D CT lesions," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 23, 2022, doi: 10.1109/TNNLS.2022.3203412.

[24] H. Liu et al., "Multi-model ensemble learning architecture based on 3D CNN for lung nodule malignancy suspiciousness classification," *J. Digit. Imag.*, vol. 33, no. 5, pp. 1242–1256, Oct. 2020.

[25] X. Liu, F. Hou, H. Qin, and A. Hao, "Multi-view multi-scale CNNs for lung nodule type classification from CT images," *Pattern Recognit.*, vol. 77, pp. 262–275, May 2018.

[26] F. Ciompi et al., "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, Dec. 2015.

[27] H. Jiang, H. Ma, W. Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1227–1237, Jul. 2018.

[28] R. Roy, S. Mazumdar, and A. S. Chowdhury, "ADGAN: Attribute-driven generative adversarial network for synthesis and multiclass classification of pulmonary nodules," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 19, 2022, doi: 10.1109/TNNLS.2022.3190331.

[29] C. Tong et al., "Pulmonary nodule classification based on heterogeneous features learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 574–581, Feb. 2021.

[30] S. Wang et al., "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Med. Image Anal.*, vol. 40, pp. 172–183, Aug. 2017.

[31] Y. Xie, J. Zhang, and Y. Xia, "Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT," *Med. Image Anal.*, vol. 57, pp. 237–248, Oct. 2019.

[32] L. Liu, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Multi-task deep model with margin ranking loss for lung nodule analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 718–728, Mar. 2020.

[33] F. E. Fernandes and G. G. Yen, "Automatic searching and pruning of deep neural networks for medical imaging diagnostic," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5664–5674, Dec. 2021.

[34] L. Yang, Q. Song, Y. Wu, and M. Hu, "Attention inspiring receptive-fields network for learning invariant representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1744–1755, Jun. 2019.

[35] Z. Lu et al., "The classification of gliomas based on a pyramid dilated convolution resnet model," *Pattern Recognit. Lett.*, vol. 133, pp. 173–179, May 2020.

[36] Y.-S. Huang, P.-R. Chou, H.-M. Chen, Y.-C. Chang, and R.-F. Chang, "One-stage pulmonary nodule detection using 3-D DCNN with feature fusion and attention mechanism in CT image," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106786.

[37] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognit.*, vol. 85, pp. 109–119, Jan. 2019.

[38] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[39] D. Zhao, Y. Liu, H. Yin, and Z. Wang, "An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118672.

[40] M. Gridach, "PyDiNet: Pyramid dilated network for medical image segmentation," *Neural Netw.*, vol. 140, pp. 274–281, Aug. 2021.

[41] L. Liu, F.-X. Wu, Y.-P. Wang, and J. Wang, "Multi-receptive-field CNN for semantic segmentation of medical images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3215–3225, Nov. 2020.

[42] Y. Wu et al., "D-former: A U-shaped dilated transformer for 3D medical image segmentation," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1931–1944, Jan. 2023.

[43] Y. Liu, Y. Zhu, Y. Xin, Y. Zhang, D. Yang, and T. Xu, "MESTrans: Multi-scale embedding spatial transformer for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 233, May 2023, Art. no. 107493.

[44] F. Chen, H. Han, P. Wan, H. Liao, C. Liu, and D. Zhang, "Joint segmentation and differential diagnosis of thyroid nodule in contrast-enhanced ultrasound images," *IEEE Trans. Biomed. Eng.*, early access, Apr. 4, 2023, doi: 10.1109/TBME.2023.3262842.

[45] S. Bhandary et al., "3D-OOCS: Learning prostate segmentation with inductive Bias," in *Proc. AAAI Conf. Artif. Intell.*, 2022.

[46] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning," 2017, *arXiv:1711.05225*.

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[49] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*.

[50] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[51] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[52] Z. Tang et al., "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[53] R. Gu et al., "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.

[54] Z. Zhang et al., "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 236–245, May 2019.

[55] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3549–3557.

[56] P. Tschandl, C. Rinner, Z. Apalla, and G. Argenzianoet, "Human-computer collaboration for skin cancer recognition," *Nature Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.

[57] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert Syst. Appl.*, vol. 128, pp. 84–95, Aug. 2018.

[58] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1938–1952.

[59] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 13, 2021, pp. 11396–11404.

[60] J. Wang et al., "Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102687.

[61] A. Khakzar et al., "Explaining COVID-19 and thoracic pathology model predictions by identifying informative input features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*, 2021, pp. 391–401.

[62] K. Tang et al., "Decision fusion networks for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 11, 2022, doi: 10.1109/TNNLS.2022.3196129.

[63] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*.

[64] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," 2016, *arXiv:1612.00410*.

[65] S. G. Armato et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, Jan. 2011.

[66] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," 2019, *arXiv:1912.12142*.

[67] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal, "A combination of shape and texture features for classification of pulmonary nodules in lung CT images," *J. Digit. Imag.*, vol. 29, no. 4, pp. 466–475, Aug. 2016.

[68] Y. Xie, J. Zhang, S. Liu, W. Cai, and Y. Xia, "Lung Nodule classification by jointly using visual descriptors and deep features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2017, pp. 116–125.

[69] A. Nibali, H. Zhen, and D. Wollersheim, "Pulmonary nodule classification with deep residual networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, pp. 1799–1808, Oct. 2017.

[70] M. Al-Shabi, B. L. Lan, W. Y. Chan, K.-H. Ng, and M. Tan, "Lung nodule classification using deep local–global networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 10, pp. 1815–1819, Oct. 2019.

[71] A. H. Chehade, N. Abdallah, J.-M. Marion, M. Oueidat, and P. Chauvet, "Lung and colon cancer classification using medical imaging: A feature engineering approach," *Phys. Eng. Sci. Med.*, vol. 45, no. 3, pp. 729–746, Sep. 2022.

Haixing Zhu received the B.S. and M.E. degrees in mechanical engineering from the Tianjin University of Technology, Tianjin, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin.

His research interests include medical image analysis.

Weipeng Liu (Member, IEEE) received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from the Hebei University of Technology, Tianjin, China, in 2010 and 2016, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Technology. His current research interests include image processing, artificial intelligence, robotics, and pattern recognition.

Zhifan Gao (Member, IEEE) received the B.S. and M.E. degrees in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2011, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He was a Post-Doctoral Fellow with Western University, London, ON, Canada, from 2018 to 2020. His research focuses on medical image processing, computer vision, and machine learning.

Heye Zhang received the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2007.

He is currently a Full Professor with the School of Biomedical Engineering, Sun Yat-sen University, Guangzhou, China. His research focuses on health informatics computing and machine learning.