Multiple Instance Learning Research for the Classification of Lung Cancer in CT Diagnosis

Huilong Chen, Xiaoxia Zhang, Tong Zhou

***Abstract*—Lung cancer with the highest mortality rate has attracted public attention. For the difficulty of treating lung cancer increases sharply over time, detecting lung cancer symptoms early on chest computed tomography (CT) is crucial for the subsequent treatment. The number of slices can affect the accuracy of lung cancer examination, so a deep multiple instance learning algorithm was designed and proposed to classify lung cancer effectively. First, feature information is extracted in the patient 3D CT image using the high and low frequency high dimensional features (HLFHD) to balance local detail and global overall information of images. Secondly, to find the decisive features, a sliding recurrent neural network (MSRNN) module is used to take into account the feature variations between slices. The experimental studies in this paper were constructed on two public datasets, namely, CIA and CC-CCII data. Finally, the experimental results show that the proposed algorithm can achieve an ACC of 0.97 and an AUC of 0.99 on the datasets. These results suggest that the proposed algorithm is well suited for lung cancer classification of any number of CT slices, and it can be effectively employed in computer-aided systems to achieve state-of-the-art performance.**

***Index Terms*—Lung Cancer; Multiple instance learning; CT images; Convolutional neural network; Deep learning**

I. Introduction

According to the World Health Organization (WHO), lung cancer is the leading cause of cancer-related deaths worldwide, accounting for the highest mortality rate among men and women. It is also important to note that there are limited treatment options for advanced lung cancer, and screening high-risk individuals has the potential to improve their survival rate by early detection. Lung carcinoma is a cancerous neoplasm that arises from the mucosal lining or glands of the bronchi. People are dying of lung cancer-related complications every day. To prevent the critical phase of tumor progression, early detection is essential for initiating treatment. Currently, CT scans are widely utilized to identify the areas affected by tumors. In clinical settings, CT images are employed for both visual and semi-quantitative assessments [1]. Beyond serving as mere images, CT scans encapsulate numerous features at the site of the lesion. These features, not readily quantifiable or assessable, requires extraction and analysis from the images to evaluate their significance [2]. Although the effectiveness of lung cancer treatment depends on the advancements in current treatment methods, early detection remains the most effective strategy for reducing mortality risk. Studies have shown that lung cancer screening using low-dose computed tomography (CT) can successfully detect disease at an early stage [3]. In the past few years, chest imaging has greatly benefited from the advancement in artificial intelligence and deep learning, becoming a specialized area of expertise. These technologies extracts features that are invisible to the human eye from various angles, including the most commonly observed histogram features, texture features, shape features, among others. The variability in the number of features that can be extracted often results in high dimensionality [4]. To retain a limited set of related features, machine learning algorithms are commonly employed for dimensionality reduction [5]. In recent years, a variety of architectures have been introduced to identify and detect specific features, aiming to address certain limitations of standard Convolutional Neural Networks (CNNs). These include Residual Networks (ResNet), Inception networks, and Dense Networks, all of which have demonstrated the ability to learn target features across diverse CT images with varying parameters [6]. However, current methods still exhibit flaws or potential issues, significantly limiting their clinical application. These models necessitate a high level of doctor involvement.

Currently, datasets pertaining to lung cancer are available from various imaging modalities, including Computed Tomography (CT), Positron Emission Tomography (PET), and X-ray. PET/CT, in particular, has been recognized as a standard imaging technique for evaluating lung cancer patients. Lung cancer is mainly composed of two types: non-small cell carcinoma and small cell carcinoma [7]. Fig. 1 illustrates the comparison between normal lung sections and various types of cancer. Medical experts believe that examining a large number of CT images of patients can mitigate the risk. However, CT scan images contain extensive nodule information. As the number of images increases, accurate evaluation becomes a challenging task for doctors [8]. Shen et al. [9] proposed a multi-scale convolutional neural network (MCNN). This network captures the heterogeneity of nodules by extracting features from stacked layers, learns features of related classes while activating the last layer of neurons, and then utilizes random forest classification to process deep features. The final accuracy achieved is 86%. Xie et al. [10] proposed a nodule classification algorithm that focuses on processing texture,

Manuscript received January 15, 2024; revised April 22, 2024.

H. L. Chen is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, 114051, China (e-mail: changango12138@163.com).

X. X. Zhang is a Professor of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, 114051, China (corresponding author, phone:86-0412-5929812; e-mail: aszhangxx@163.com).

T. Zhou is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, 114051, China (e-mail: 3223464820@qq.com).

shape, and deep information to classify each nodule. The final accuracy achieved by the algorithm is 96%. Ardila et al. [11] proposed an end-to-end 3D CNN model to calculate the overall risk of lung malignant tumors using the fully public NLST dataset. Compared with six radiologists, the model reduced false positives by 11% and false negatives by 5%. The model's performance is comparable to that of radiologists. Gopinath et al. [12] proposed a neural network model that combines GAN and CNN. They utilized grayscale conversion, scaling, and denoising as preprocessing steps for images. The CNN model was trained using the watershed threshold method and enhanced GAN-mask technology to classify lung cancer. Xiao et al. [13] introduced a multiple feature multiple attention network (MFMANet), incorporating a multiple scale spatial attention module (MSAM) and a multiple feature fusion attention module (MFGLA) to improve the detection of small lesion areas. Their model achieved an accuracy of 99.06% and 91.67% on datasets of lung adenocarcinoma and lung squamous cell carcinoma, respectively. Through multiple instance learning (MIL) based on attention mechanisms, the system can handle packets of different sizes and effectively express the distribution of key features within a packet [14], resulting in high visibility of features [15]. In multiple instance learning, convolutional neural networks (CNNs) are also employed as feature extraction methods. However, CNNs with insufficient depth may fail to extract effective features, which is contingent upon the complexity of the problem. Ilse et al. [16] utilized a two-layer convolutional neural network to extract effective features from the MNIST-based MIL dataset. It's noteworthy that the resolution of the images in this problem (512×512) is significantly larger than that of the MNIST dataset images (28×28). Selecting a deeper and wider ResNet [7] or EfficientNet [17] can effectively address this challenge. Another issue to consider is the impact of the network's width on the overall neural network performance. Selecting an appropriate width enhances the model's convergence and reduces computational requirements. In other words, the convergence of convolutional neural networks (CNNs) can be improved by increasing the width of the deep neural network (DNN) [18]. Alakwaa et al. [19] introduced a classification model based on 3D CNN, employing a 3D convolutional neural network to preserve the spatial structure information within CT volumes and comprehend the alterations in cancerous regions within the images.

The prevailing lung cancer classification algorithms aim to accomplish precise and swift categorization of cancerous image cases. Despite their commendable performance in cancer image processing applications, these methods exhibit several shortcomings. Many models lack sufficient feature extraction capabilities, leading to suboptimal quality of

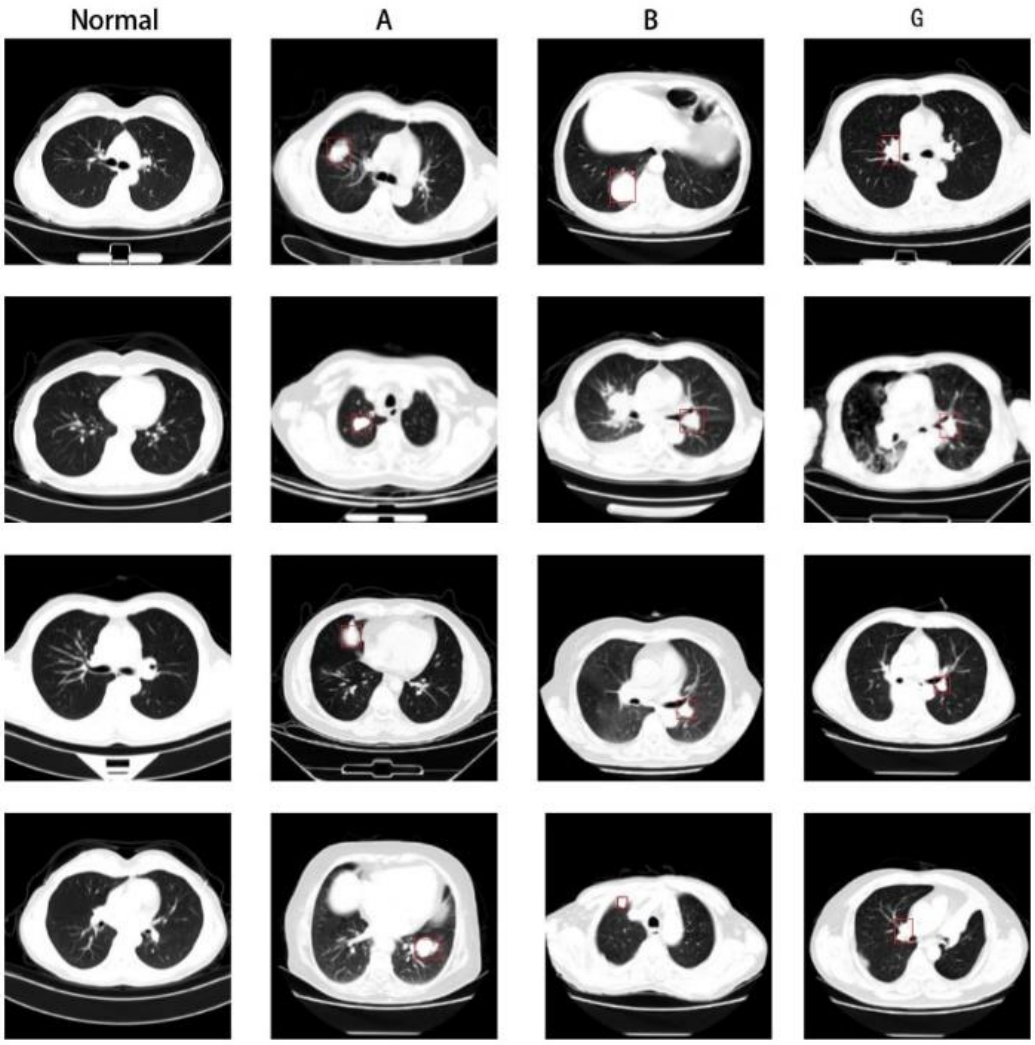

Fig. 1. Lung cancer type (Normal: normal, A: adenocarcinoma, B: small cell carcinoma, G: squamous cell carcinoma)

crucial information within the images. Moreover, in real-world scenarios, delineating specific cancerous areas in each CT slice proves to be expensive and time-intensive for medical professionals. Therefore, a method that only necessitates marking whether the entire 3D CT volume contains cancerous regions becomes particularly crucial. Furthermore, such a method should be capable of handling 3D CT volumes of varying sizes.

Based on the current research status, we propose a lung cancer classification model grounded in deep multiple instance learning, with the objective of achieving precise and interpretable screening of lung cancer cases using chest CT scans. Adhering to the definition of multiple instance learning, each case's 3D CT volume is conceptualized as a bag, while each slice within it is considered an instance. The primary aim of our proposed model is to learn a single label for the predicted case bag and to generate deep instance features with permutation invariance. Through experimental evaluation on two public datasets, our results demonstrate the superior performance of the proposed model compared to existing lung cancer classification methods. The paper's main contributions are summarized below:

(1) In this work, we employ extensive preprocessing techniques to extract key features from lung images, aiming to minimize noise and enhance the accuracy of lung cancer classification. Additionally, we train the model end-to-end from scratch, ensuring comprehensive learning and optimization throughout the process.

(2) In this paper, we introduce a novel module called the High-Low-Frequency High Dimensional Feature Extraction (HLFHD-RESNET) module, designed to simultaneously capture high-frequency and low-frequency features. High-frequency features encapsulate image details and edges, characterized by rapidly changing pixel values. Conversely, low-frequency features depict overall structure and large-scale changes, characterized by slowly changing pixel values. These features are combined into a 2D feature map, enhancing the model's generalization capability, enabling adaptation to various data distributions, and mitigating the impact of inherent noise in the dataset on the learning process.

(3) In this paper, we introduce a sliding recurrent neural network (MSRNN) module designed to forecast the probability of a fixed-length sequence within the 2D feature information graph. Subsequently, we derive the weight distribution of packet-level features by combining the obtained probability graph with the attention mechanism. Finally, the classifier determines the category based on these derived weights.

(4) This paper introduces a model based on multiple instance learning, featuring an "end-to-end" network architecture. Through experimental evaluation on two extensive public datasets, TCIA and CC-CCII, we demonstrate that the model exhibits superior robustness and generalization capabilities.

II. Related Works

In recent years, deep neural networks have demonstrated remarkable achievements in various computer vision tasks, showcasing immense potential in image feature learning. By augmenting the depth and width of the network, researchers aim to capture increasingly complex and abstract feature representations, thereby facilitating the completion of tasks through the utilization of retained relevant information. In response to diverse task requirements and objectives, researchers in this field continuously refine the network structures and develop classification algorithms with enhanced performance and generalization capabilities, aiding medical professionals in making accurate diagnoses. In alignment with the objectives of this paper, this section provides an overview of related work in lung cancer classification and outlines the designed methodologies employed in previous studies.

A. Dataset preprocessing

A total of 254 lung cancer patients (191A, 29B, 34G) and 243 disease-free participants were enrolled in this study. The patient image data were obtained from TCIA. Prior to the examination, each patient fasted for at least 6 hours, and their blood glucose level was maintained below 11 mmol/L. Whole-body emission scans were conducted 12 minutes after intravenous injection of 60F-FDG (18.4MBq/kg, 44.0mCi/kg). The disease-free data were sourced from the China Chest CT Imaging Research Consortium (CC-CCII).

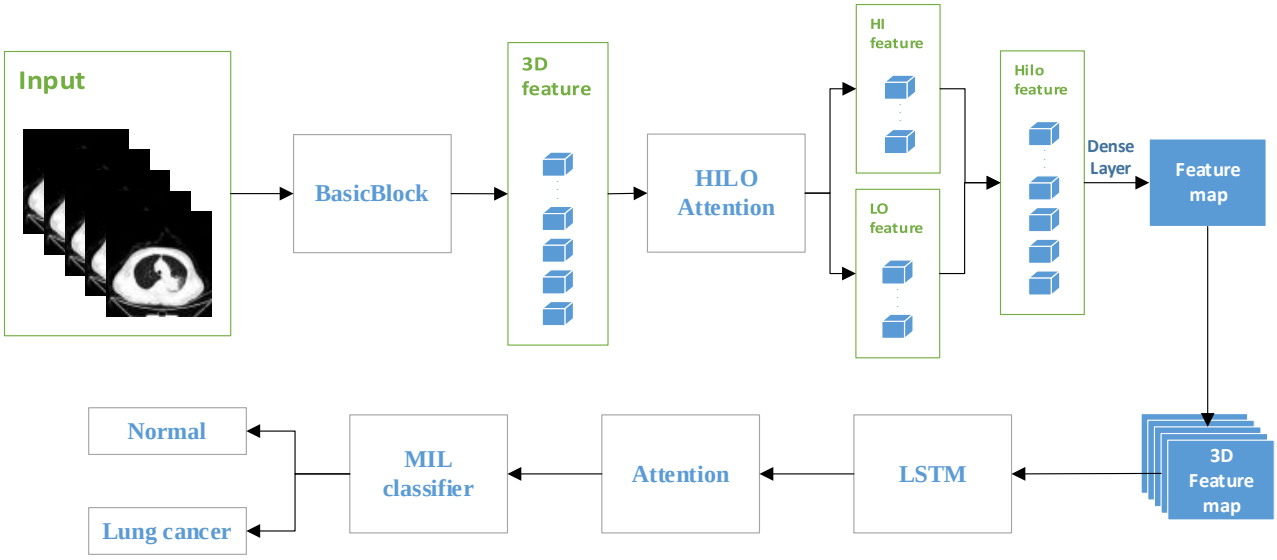

Fig. 2. Overview of the HLFHD-RESNET framework.

Since the two parts of data come from different places and the file formats are different, this paper will unify the CT imaging conditions of the two parts of data. The original CT image is composed of anisotropic voxels with different in-plane resolutions. Due to different scanners or different acquisition protocols, CT datasets with different voxel spacing are generated. In order to facilitate training, all medical images are resampled according to the voxel spacing provided by the DICOM file, and then the resolution is uniformly adjusted to be consistent. The slice thickness ranged from 0.625 mm to 5 mm, and the scanning modes included plain weave, contrast and 3D reconstruction.

The images were analyzed using a window width of 1050 HU and a window level of -475 HU. Reconstruction was conducted with a section thickness of 2 mm within a lung environment. The CT slice interval ranged from 0.625 mm to 5 mm, with scanning modes including plain weave, contrast, and 3D reconstruction. The location of each tumor was annotated by five academic chest radiologists specializing in lung cancer. Two radiologists possessed over 15 years of experience, while the remaining three had more than 5 years of experience. Following annotation by one radiologist, the other four radiologists performed verification. Additionally, each of the five radiologists reviewed every annotation file in the dataset.

While it is recognized that CC-CCII solely offers CT images in JPEG or PNG format, it is acknowledged that the compression inherent to JPEG from DICOM may potentially impact clinical diagnosis performance. Nevertheless, through thorough comparison and debugging efforts, it has been verified that the data from both sources are ultimately displayed under uniform conditions, thereby ensuring that training is unaffected. Consequently, the obtained results retain significant application value.

B. Preparing a slice-based instance

Given the diverse sources of CT images, priority is assigned to processing data from TCIA. To align with the CT value parameters (1050 HU, -475 HU) utilized for extracting lung features, the window width and window level of CC-CCII data are adjusted accordingly. Fig. 1 illustrates a well-processed data sample. Subsequently, the data is normalized, with each pixel value scaled to fall within the range of 0 to 1. Following normalization, each CT image undergoes resampling, with voxel spacing ranging from 0.585937 to 0.841796. The resolution range after resampling lies between 300 and 431. Next, the images are resized to dimensions of 256 × 256 using the OpenCV tool. Finally, each processed image is encapsulated into a bag.

C. Extraction of deep features

In this paper, the fine-tuned ResNet-50 and Hilo attention [20] serve as the primary network for deep feature extraction (Fig. 3), with ResNet-50 focusing on capturing hierarchical features and Hilo attention emphasizing important regions within the images. The final fully connected layer of the network is responsible for outputting features, with the output size matching the number of channels in the preceding layer to preserve feature information. The network parameters are initialized randomly. The initial learning rate is set to 0.0005, and it is subsequently decayed using the cosine function to ensure stable training. A total of 50 training epochs are conducted. It's worth noting that networks with insufficient depth may struggle to abstract image features effectively and fail to concentrate on key feature distributions, even when sufficient data is available, although they may suffice for simpler computer vision tasks [21]. As the problem complexity escalates, the network depth tends to increase accordingly, and the quality of feature expression becomes increasingly reliant on the training dataset. The length of a packet is set to n, and upon inputting a packet into the network, a ($n\times512$) feature map is generated, resulting in a total of 512 packet-level depth features, which are then used for further processing and classification tasks.

D. Recurrent Neural Network Based Inference

Initially, it's essential to acknowledge that CT images constitute 3D-level data, wherein the slices arranged sequentially retain significant spatial information. However, the approach outlined in the previous method [22] disregards this spatial information and solely extracts 2D-level features from each slice. Subsequently, these features are combined via pooling to perform feature mapping. Despite considering all slices, the features are extracted independently, thus lacking correlation between slices within the feature information.

Many deep learning models leverage 3D CT images as input [23], necessitating a preprocessing step to choose a fixed number of slices for model input. For instance, S et al. [24] opted for a fixed number of slices and introduced the Cloud-YLung model for histological classification of NSCLC based directly on 3D CT images from whole lung scans. In this process, a crucial consideration is how to select 3D CT images of consistent size as input across different packet sizes. Given variations in patient requirements, imaging at different slice intervals might result in the disappearance or attenuation of lesions if the CT image packet's slice interval is altered. Alternatively, manual slice selection can be employed, but this approach risks overlooking potentially affected slices and increases the workload for doctors.

In summary, this paper addresses two key issues: the variable number of slices and the spatial relationships between slices. To tackle these challenges, we propose an MSRNN framework designed to learn the spatial relationships between slices through a reasoning process. Additionally, the framework incorporates an attention mechanism to handle varying numbers of CT image packets.

E. Multiple instance learning

Methods based on multiple instance learning (MIL) [25] play an important role in addressing the aforementioned problems. In this paper, all CT images of the patient are referred to as instances, which can be either lung cancer positive or negative, and the model is trained under weakly supervised conditions. Most MIL-based methods draw inspiration from Ilse et al. [16], who proposed an attention mechanism to learn the correlation confidence between different instances to assess patient-level classification. Shi et al. [26] mentioned that the prediction at the packet level largely depends on the validity of the learned instance weights. To address this, a loss-based attention mechanism

is proposed, where instance weights are calculated using the SoftMax + cross-entropy loss function, and the parameters are shared with the fully connected layer for predicting instances and packets.

In the model proposed in this paper, HLFHD-RESNET is utilized for integrating high and low frequency features in the first stage of feature extraction. Here, the first stage is considered as generating a feature map. In the second stage, SRNN is employed to process high-dimensional features among instances and learn the relationships between them. Finally, effective relationship information is obtained through the attention mechanism to predict the packet.

III. Methods

In the current study of multiple instance learning, the dataset unit is referred to as a "bag," which comprises multiple instances. When all instances within a bag are labeled negative, the bag is termed a negative bag, and conversely, it is called a positive bag. The ratio of positive to negative instances in a positive bag significantly influences each instance's contribution ratio. However, most algorithms assume that both positive and negative instances are independently sampled from their respective distributions, which often does not meet the requirements of practical problems. This is due to the structural and interrelationship characteristics between bags and instances. For instance, instances may be sequential or exhibit spatial and temporal dependencies. The randomness inherent in sampling poses significant challenges in capturing packet-level features. To address these issues, it is imperative to enhance both the feature extraction method and the MIL classification method. In this section, we introduce the methodological flow of each module of the proposed model based on the aforementioned research direction. The subsequent sections provide detailed descriptions of each module and the developed model.

A. Overall Structure of the Network

In this paper, we introduce a novel image classification model. The network model integrates a residual learning module, a high and low frequency attention mechanism, and a recurrent neural network. Fig. 4 illustrates the overall structure of the network model, which effectively extracts feature information and achieves accurate classification.

The figure illustrates the input data to the model, which consists of a 3D CT volume. The bottom of the figure represents the dimension of the data after passing through each module. The model utilizes the residual learning module to perform down sampling operations, gradually reducing the size of the input to enhance the perception of the convolution kernel. The Hilo attention mechanism is utilized to enable the model to focus on high-frequency and low-frequency information in the image separately, and the newly designed HLFHD-RESNET module is integrated by connection. This module incorporates a multi-head attention mechanism and pooling technology to acquire high and low frequency features of the image. In the output, the maximum number of channels is constrained to 512 to minimize computational complexity. The detailed process of the proposed model includes the following steps:

- Step 1: Input image preprocessing involves preprocessing the original CT image, including adjusting image resolution, voxel spacing, and resampling, to ensure that the data meets the requirements of model training.
- Step 2: HLFHD-RESNET module. The preprocessed CT image is fed into the module, where high-level feature representations are obtained through a series of convolutional layers and pooling layers within the residual block. Subsequently, the feature information from different perspectives is captured by the multi-head attention mechanism. This ensemble of feature information constitutes the integration of global high-frequency attention and local low-frequency attention.
- Step 3: MSRNN module. The feature map is rearranged based on the window size, followed by the application of a recurrent neural network to identify the target part within the new sequence, namely, the lesion probability distribution. Subsequently, the attention mechanism is utilized to map the probability distribution onto the instance sequence.
- Step 4: MIL classifier. The probability sequence is fed into the MIL classifier to predict the packet label. This classifier is capable of learning various levels of semantic and contextual information within the image, thereby enabling more accurate predictions.

B. HLFHD-RESNET for 2D level diagnostics

In this section, the HLFHD-RESNET module is constructed to develop features that can be extracted at the application level. This module incorporates a Hilo attention layer [20] before global average pooling, enabling the extraction of high and low-frequency features from the feature maps obtained from each convolutional kernel. In Fig. 2, the high-frequency attention component aims to capture the dependencies of fine local features, encoding the local details of the object. It sets the local self-attention window to capture fine feature information and utilizes non-overlapping window partitioning to reduce redundant time-consuming operations. Conversely, the low-frequency

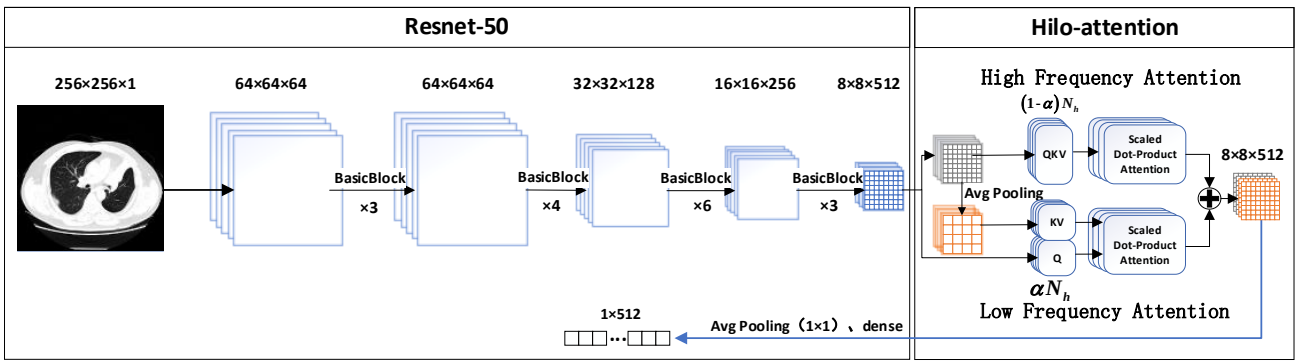

Fig. 3. Overview of the HLFHD-RESNET framework

attention component is designed to capture rich low-frequency information within the features. The low-frequency attention component conducts average pooling for each window to obtain low-frequency signals in the input. Subsequently, it maps these low-frequency features to Key and Value keys, with Query as input. Finally, the feature map for the next input is obtained by concatenating the high-frequency information with the low frequency information and then calculating its global average pooling with the window set to (1×1), followed by passing through the fully connected layer.

C. *Multiple instance learning*

According to the standard MIL formula, each patient's CT image is regarded as a bag with each slice considered as an "instance." The label of the bag is associated with the corresponding instance label, implying that all instances within the same bag share the same label. However, due to the fact that only a very small number of instances may actually satisfy the label [27], while most other instances may be contrary to the label, this unbalanced allocation method may introduce noise to the positive bag. Consequently, adjustments to the model parameters are necessary to enable better differentiation between different categories or to perform more accurate discrimination tasks.

In the multiple instance problem, let $X=\{x_1,x_2,...,x_K\}$ denote a set of K instances in a bag. Given different bags, the size of k may vary. Each bag corresponds to a label $Y\in\{0,1\}$, and it is assumed that each instance in the bag has an individual label associated with it, denoted as $y_1,y_2,...,y_k$, and $y_k\in\{0,1\}$. However, the essence of the regression problem is that the instance label cannot be accessed during the training process. Thus, the MIL problem hypothesis [16] can be expressed as follows:

$$Y=\begin{cases}0, & iff\ \sum_k y_k=0\\ 1, & otherwise\end{cases} \tag{1}$$

In general, a prediction model for the MIL problem requires two functions: one is the appropriate transformation function, and the other is the permutation invariant function. The prediction function of the model is defined as follows:

$$P(X)=g\left(\sum_{x\in X} f(X)\right) \tag{2}$$

For given f and g, there are two main MIL approaches:

(1) Based on the instance-based approach, f is used as an instance classifier. The transformation function f is considered as an instance classifier, and scores are calculated for each instance. g is considered as an identity function of the pooling operation type. The prediction probability of the bag is obtained by summarizing the scores of each instance.

(2) Based on the embedding approach, f is used as a feature extractor to embed the features of each instance in low dimensions respectively. g is considered as an aggregation operator. It aggregates the low-dimensional embedding of the instance into a bag-level embedding and generates the prediction probability of the bag according to the embedding of the bag.

In these two methods, [22,25,28] argue that the embedding-based method is superior to the instance-based method in all aspects. Since the label of the instance is unknown during the training process, and the instance classifier may be affected by the imbalance in the number of instances corresponding to each label, leading to errors in the final prediction. Models capable of identifying key instances can make better predictions for bag labels. Consequently, the article adopts an embedded perspective on the method.

D. *MIL with HLFHD-RESNET module*

In the classical MIL problem, the instance does not require further processing of the representation results of the features and is considered to be an identity. However, when dealing with other complex problems such as images or texts, it becomes necessary to enhance the feature extraction method; otherwise, it would be impossible to generate features that determine the prediction results. A feature extraction method that combines residual neural network [6] and Hilo attention [20] as a model is proposed. As shown in Fig. 2, let M be the output vector of the residual block, then:

$$M(x_k)=F(x_k,\{W_i\})+x_k \tag{3}$$

where $x_K\,(x\in X)$ is the input vector of the residual block, and the function $F(x_K,\{W_i\})+x_K$ represents the base mapping of the residual block. It is obtained by summing with the initial input A to produce an output vector that fully encompasses the range of the previous residual block. The Kth instance is transformed into a low-dimensional embedding through multiple residual blocks.

When the embedded features of the respective instances are obtained, the subsequent step involves acquiring the high and low-frequency feature vectors within the features obtained across various instances. Firstly, a multi-head self-attention mechanism (MSA) [29] is established, which is capable of capturing relationships between different positions. Let $M\in R^{N\times D}$ denote the input, where N is the length of the input vector, D is the size of the hidden dimension, and each self-attention headsets Q (Query

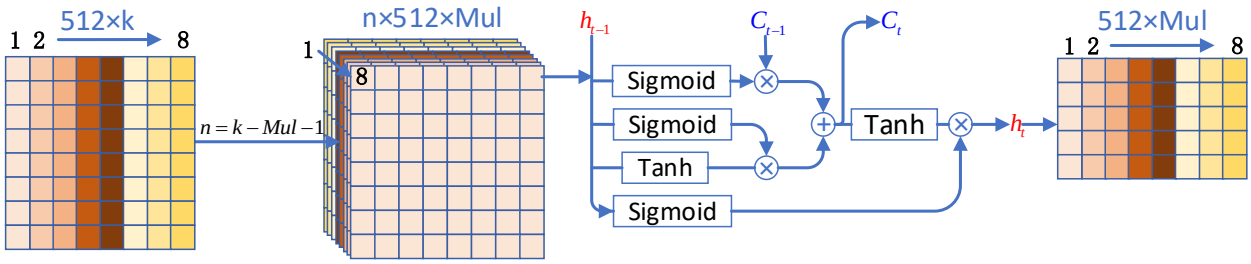

Fig. 4. Overview of the MSRNN framework.

matrices), K (Key matrices), and V (Value matrices):

$$Q = m_k W_q^h, \quad K = m_k W_k^h, \quad V = m_k W_v^h \tag{4}$$

Where $W_q^h, W_k^h, W_v^h \in R^{D \times D_h}$ is the number of self-attentive heads, D_h denotes the number of hidden dimensions of the h-th head. Then, each head outputs a weighted sum of input vectors.

$$SA_h(m_k) = Softmax\left(\frac{Q_h K_h^T}{\sqrt{D_h}}\right)V \tag{5}$$

Then the input of each head is combined:

$$MSA(m_k) = concat\left[SA_h(m_k)\right]W_o \tag{6}$$

Where $W_o \in R^{(N_h \times D_h) \times D}$ is a weight matrix acquired through learning, which is employed to map the aggregation results from multiple attention heads to the final output.

In high and low-frequency attention mechanisms, a hyperparameter $\alpha \in [0,1]$ is defined to allocate the number of high and low-frequency attention heads, where $\alpha \times d$ represents the number of high-frequency attention heads, and $(1-\alpha) \times d$ denotes the number of low-frequency attention heads. By adjusting α, we can decide whether to prioritize high or low frequencies based on the specific problem, thus effectively allocating computational resources.

High-frequency attention (HF) encodes high-frequency features directly through local attention applied to the input. Conventionally, high-frequency features are associated with capturing local details of objects. Therefore, a local self-attention window, typically of size 2 × 2, is designed to capture these high-frequency features.

Low-frequency attention (LF) encodes low-frequency features through down-sampling the global attention of the input. In Fig 2, average pooling is applied to each window of the input to derive the low-frequency signals in M. Subsequently, the averaged-pooled feature maps are mapped to the keys K and V, while the query Q is still obtained from the mapping of M. This approach reduces the complexity of Equations (4) and (5) compared to high-frequency attention.

Then the output of each attention is spliced together:

$$Hilo(M) = concat\left[HF(M), LF(M)\right] \tag{7}$$

Finally, the output of $Hilo(M)$ will be processed through global tie pooling and the linear layer, resulting in a 2-dimensional feature map ($k \times l$), where k represents the size of the packet and l represents the size of the output of the linear layer.

E. Attention-based MSRNN module

In classical multiple instance learning (MIL) problems, such as those encountered in handwritten digit datasets, each instance within each bag is typically considered independent and unrelated [25]. However, in the case of a complete 3D CT image, the data is continuous, and it is also presumed that the features generated by different instances under similar conditions exhibit continuity.

When addressing this issue, it's possible to encounter scenarios where more than one cancer lesion exists within a bag containing cancer instances, and instances of the same cancer with morphological variations may appear in consecutive slices. To address these situations, it is proposed to employ a recurrent neural network for processing variable-length sequences. Here, the sequence features derived from all instances under similar conditions are referred to as a sequence.

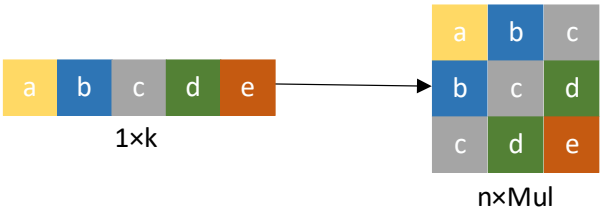

Fig. 5. Demonstration of the process of changing Mul values in the MSRNN module

First, the 2D feature map $H(k \times l)$ obtained in the previous section is transposed. Due to the varying sizes of different packets, the value of k becomes uncertain. The conventional method is to establish a fixed-length template sequence and apply zero-padding to different sequences. However, considering that the zero-padding operation can be influenced by $tanh\,function(\in[-1,1])$, a sliding window is introduced in this process to handle sequences of length k. As depicted in Fig. 5, the length of the sliding window is set to Mul, processing sequences of Mul lengths:

$$n = k - Mul - 1 \tag{8}$$

After this processing, the length of the sequence for each input recurrent neural network is standardized. It is essential to note that the input shape is $(l \times k)$, the shape of the MS function is $(n \times l \times Mul)$, and l represents the batch size entering the recurrent neural network, that is:

$$gf = MS\left(Hilo^T\right) \tag{9}$$

Then, gf is fed into the LSTM network, where each batch in gf comprises n sequences of length Mul. Hence, it is imperative to compute the gating units at each time step, referring to the mathematical formulation of LSTM [30]:

$$i_{n_t} = sigmoid\left(W_i gf_{n_t}^p + U_i h_{t-1}\right) \tag{10}$$

$$f_{n_t} = sigmoid\left(W_f gf_{n_t}^p + U_f h_{t-1}\right) \tag{11}$$

$$o_{n_t} = sigmoid\left(W_o gf_{n_t}^p + U_o h_{t-1}\right) \tag{12}$$

$$\tilde{c}_t = tanh\left(W_c gf_{n_t}^p + U_c h_{t-1}\right) \tag{13}$$

$$c_t = i_{n_t} \odot \tilde{c}_t + f_{n_t} \odot \tilde{c}_{t-1} \tag{14}$$

$$h_t = o_{n_t} \odot tanh(c_t) \tag{15}$$

Where W and U are weight matrices, $gf_{n_t}^p$ is the vector input of the time step t of the n-th sequence of the p-th batch, h_t is the hidden state at the current time step, c_t is the memory cell at the current time step, and $\odot$ denotes element-wise multiplication. Finally, splicing all inputs together will result in a two-dimensional matrix $(n \times p)$.

The significance of the final output of the MSRNN

module lies in acquiring its relational weights across the instances' lengths. However, similarity might emerge among instances that are closely situated due to the minimal difference between the first and second steps being solely the variance between the two instances. Therefore, it is proposed to establish an attentional weight for the relation matrix and employ a weighted average for the relation of each segment of the instances, with the weights determined by the neural network. Let $H=\{h_1,h_2,...,h_n\}$ denote the relationship of the n-th segment instance, as follows:

$$z=\sum_{n=1}^{n} a_n h_n \tag{16}$$

Where:

$$a_n=\frac{exp\{W^T tanh\left(Vh_n^T\right)\}}{\sum_{j=1}^{n}\exp\{W^T tanh(Vh_j^T)\}} \tag{17}$$

Where $W\in R^{L\times 1}$ and $V\in R^{L\times M}$ represent the parameters of the neural network. The validity of this construction will be experimentally verified thereafter.

IV. Experiments

A. Datasets

In this study, two of the largest publicly available lung CT datasets, TCIA [31] and CC-CCII, have been chosen for experimentation. Each package contains a varying number of slices, ranging from 30 to 110, and will be utilized to assess the proposed model. This section elaborates on these two datasets in Table 1. The TCIA dataset comprises three categories: A (adenocarcinoma), B (small cell carcinoma), and G (squamous cell carcinoma), while the CC-CCII dataset exclusively contains cases labeled as Normal (no cancer). In total, there are 36,031 slices across 600 packages, with 460 packages allocated for training and validation, and the remaining 140 packages reserved for testing. The training and validation ratio is set at 4:1.

In this experiment, all slices from the two datasets are saved as tensor format data, with each slice having a resolution of 256×256. This downsizing reduces the size to one-fourth compared to the original slices, resulting in a significant reduction in dataset file size. Furthermore, it drastically shortens training time and alleviates the demand on video memory of the graphics card. After testing, it was found that the downsized data exhibits little difference compared to the original data. The experiments will randomly allocate data into training, validation, and test sets in a 6:2:2 ratio.

In this study, two of the largest public lung CT datasets, TCIA and CC-CCII, are chosen for experimentation. Each package contains a varying number of slices, ranging from 30 to 110, which will be employed to evaluate the proposed model. Detailed information about these datasets is provided in Table 1 of this section. The TCIA dataset encompasses three categories: A (adenocarcinoma), B (small cell carcinoma), and G (squamous cell carcinoma). Conversely, the CC-CCII dataset exclusively consists of Normal (no cancer cases). In total, there are 36,031 slices distributed across 600 packages. Out of these, 460 packages are allocated for training and validation, while 140 packages are designated for testing. The ratio of training to validation is set at 4:1.

B. Experimental details

All experiments are conducted on a local workstation equipped with an Intel(R) Core(TM) i9-12900H processor and an NVIDIA GeForce RTX 3070 Ti Laptop GPU. A large number of experiments are then performed to evaluate the sensitivity of the hyperparameters. Specifically, the initial learning rate of the model is set to 0.0005 based on prior experience, and the learning rate of the optimizer is adjusted using the cosine annealing method. The learning rate starts from the initial value and gradually decreases via cosine annealing until the maximum number of iterations (T_{max}) is reached, with T_{max} set to 20. The model is trained using the Adam optimization algorithm, retaining the default parameters β_1 and β_2. The batch size is set to 1, the number of epochs is set to 50, and the dropout rate α of the HLFHD-RESNET module is set to 0.5.

The model proposed in this paper integrates research methodologies from deep learning and offers the advantage of being equally applicable to both small and large datasets. In comparison with similar methods, it possesses additional advantages, such as its weak supervision nature of learning and the avoidance of lesion segmentation requirements.

C. Evaluation metrics

The training and testing procedure of the proposed model employs 5-fold cross-validation. In this approach, 4/5 of the data is utilized for training the model, where the model undergoes fine-tuning with pre-training parameters. The remaining 1/5 of the data is reserved for validation. The performance of the model is evaluated using five standard classification performance metrics, namely, the area under the ROC curve (AUC), accuracy (ACC), sensitivity (SEN), specificity (SPE), and F1 score:

TABLE I
A DESCRIPTION OF THE LUNG CT IMAGE DATASET

Datasets	Classes	Slices		Bags	
		Train & Val	Test	Train & Val	Test
TCIA	A	9045	2683	167	49
	B	1712	386	34	7
	G	1629	741	29	14
	Total	12386	3810	230	70
CC-CCII	Normal	15318	4517	230	70

TABLE II
COMPARISON OF MODEL PERFORMANCE FOR IDENTIFYING DIFFERENT α VALUES.

α	Initial lr	AUC	ACC	SEN	SPE	F1 score	Precision
$\alpha=0.1$	0.0005	0.881020	0.800000	0.671428	0.928571	0.770491	0.903846
$\alpha=0.2$	0.0005	0.875510	0.864286	0.985714	0.742857	0.878980	0.793103
$\alpha=0.3$	0.0005	0.940816	0.878571	0.914285	0.842857	0.882758	0.853333
$\alpha=0.4$	0.0005	0.945510	0.885714	0.842857	0.928571	0.880597	0.921875
$\alpha=0.5$	0.0005	0.952245	0.907143	0.814285	1.0	0.897637	1.0
$\alpha=0.6$	0.0005	0.938367	0.900000	0.9	0.9	0.9	0.9
$\alpha=0.7$	0.0005	0.947755	0.892857	0.8	0.985714	0.88189	0.982456
$\alpha=0.8$	0.0005	0.947143	0.828571	0.742857	0.914286	0.8125	0.896552
$\alpha=0.9$	0.0005	0.943877	0.814286	0.714285	0.914285	0.793650	0.892857

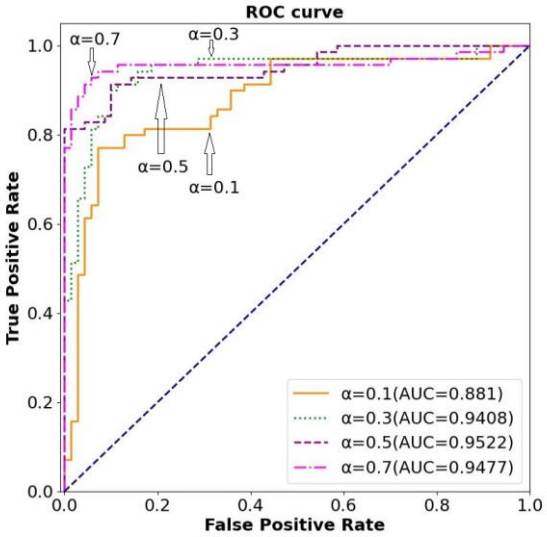

Fig. 6. In setting the ROC curves and AUC values for models with different α values

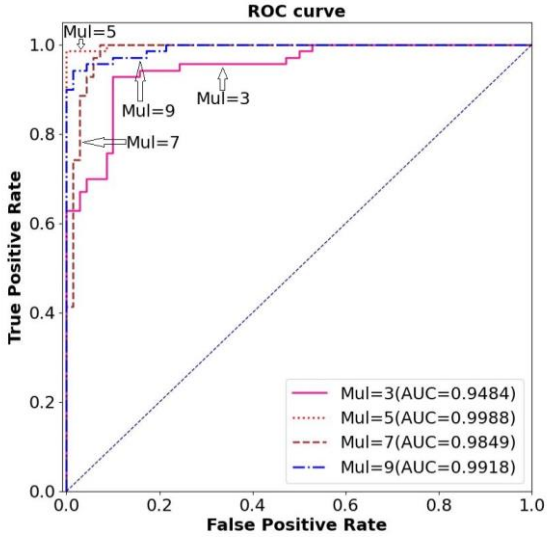

Fig. 7. In setting the ROC curves and AUC values for models with different *Mul* values

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$SEN = \frac{TP}{TP + FN} \tag{19}$$

$$SPE = \frac{TN}{TN + FP} \tag{20}$$

$$F1\,score = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{21}$$

Where TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives, respectively. Additionally, $Precision = \frac{TP}{TP + FP}$ and *Recall* are equal to *SEN*. The $F1\,score$ mitigates the

TABLE III
COMPARISON OF MODEL PERFORMANCE FOR IDENTIFYING DIFFERENT *Mul* VALUES.

Mul	Initial lr	AUC	ACC	SEN	SPE	F1 score	Precision
$Mul=2$	0.0005	0.955306	0.864286	0.785714	0.942857	0.852713	0.932203
$Mul=3$	0.0005	0.948367	0.900000	0.9	0.9	0.9	0.9
$Mul=4$	0.0005	0.927755	0.942857	1.0	0.885714	0.945946	0.897436
$Mul=5$	0.0005	0.998776	0.971429	0.942857	1.0	0.970588	1.0
$Mul=6$	0.0005	0.983878	0.964286	1.0	0.928571	0.965517	0.933333
$Mul=7$	0.0005	0.984898	0.942857	1.0	0.885714	0.945946	0.897436
$Mul=8$	0.0005	0.994694	0.928571	1.0	0.857143	0.933333	0.875
$Mul=9$	0.0005	0.991837	0.914286	0.828571	1.0	0.90625	1.0
$Mul=10$	0.0005	0.91	0.885714	0.842857	0.928571	0.880597	0.921875

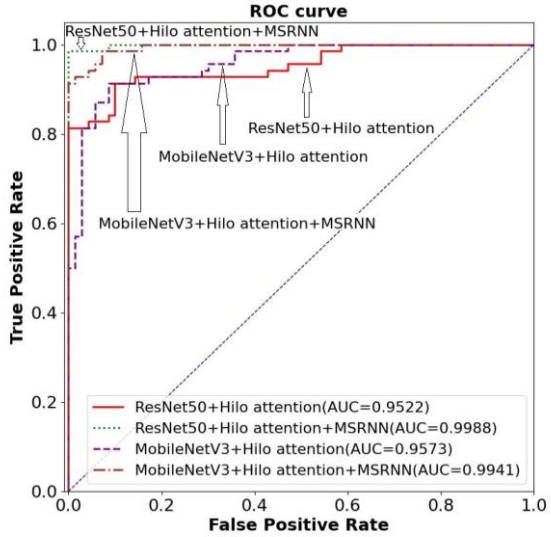

Fig. 8. ROC curves and AUC values of the proposed model for different scenarios

interference of unbalanced data.

Next, the ablation experiment will be conducted to verify each module of the model. In this paper, the HLFHD-RESNET module and the MSRNN module are integrated. The performance of both modules will be evaluated based on their respective metrics and significance. Detailed evaluation indicators for the two modules will be provided below.

D. Experiments based on the HLFHD-RESNET module

In this section, we will conduct experiments on the hyperparameter α of the high and low frequency attention in the HLFHD-RESNET module, denoted as $\alpha \in \{0.1, 0.2, ..., 0.9\}$. The α value will be varied across 9 values during the experiment to train the model. Specifically, the model selection for this experiment comprises ResNet50 and HLFHD-RESNET, excluding the MSRNN module. Afterwards, an appropriate value of α will be selected based on evaluation metrics for use in subsequent experiments.

By comparing the performance of the model under different values set in Table 2, it was found that when $\alpha = 0.5$, the model exhibits the highest performance level, with an accuracy (ACC) of 0.9071, area under the ROC curve (AUC) of 0.9522, and F1 score of 0.8976. Therefore, the settings of $\alpha = 0.5$ will be maintained in subsequent experiments. Fig. 6 illustrates the area under the ROC curve for different settings of α.

E. Experiments based on the MSRNN module

In this section, we will experiment with the size of the hyperparameter *Mul* of the MSRNN module, denoted as $Mul \in \{2, 3, ..., 10\}$. The experiment involves setting *Mul* to 9 different values for training the model, and subsequently comparing the evaluation indicators of the model. For this experiment, the model utilizes ResNet50, with the hidden layer state dimension of the LSTM set to 1, and the number of LSTM layers set to 1.

By comparing the different sizes of the *Mul* values of the MSRNN module in Table 3, it is evident that the model achieves its best performance when $Mul = 5$, with an accuracy (ACC) of 0.9714, area under the ROC curve (AUC) of 0.9987, and F1 score of 0.97. Therefore, we will continue the experiment with this setting in subsequent experiments. Figure 7 illustrates the area under the ROC curve for different settings of *Mul*.

F. Model Performance Evaluation

In this section, the experiments will be divided into six

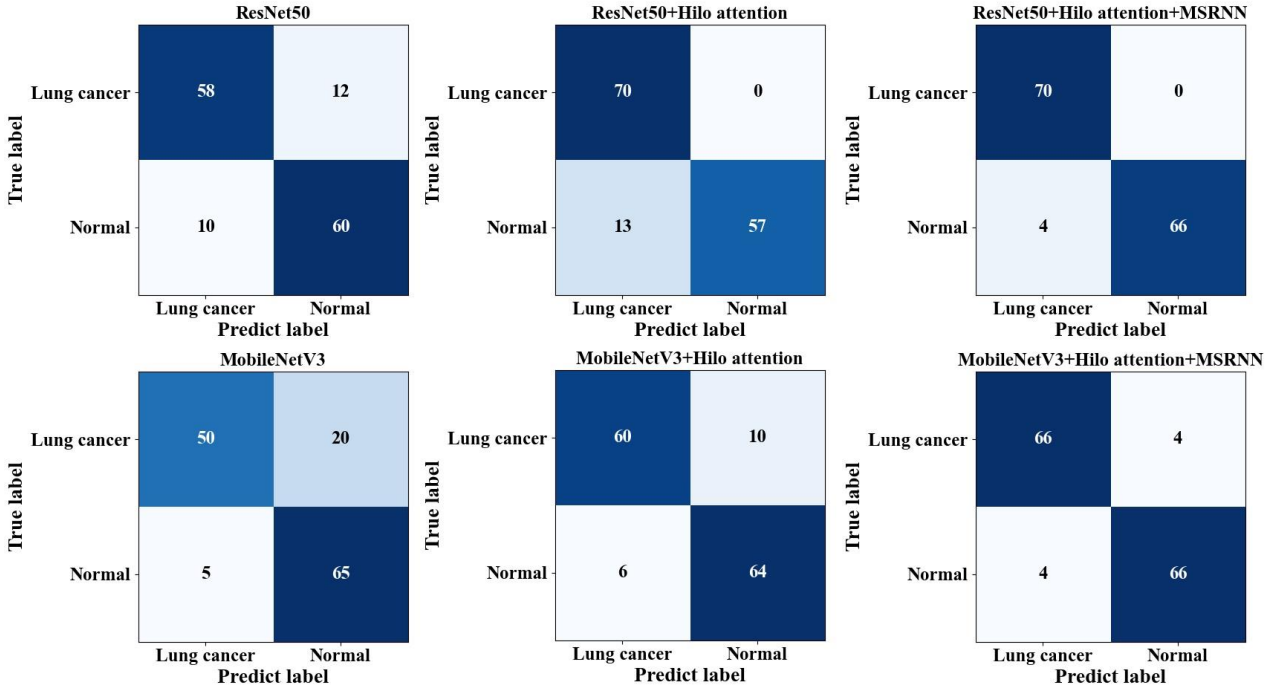

Fig. 9. Confusion matrix for the proposed and comparable models

TABLE IV
COMPARISON OF MODEL PERFORMANCE FOR SEVERAL DIFFERENT MODULE COMBINATIONS

Model	AUC	ACC	SEN	SPE	F1 score	Precision
ResNet50	0.914897	0.842857	0.885714	0.8	0.849315	0.815789
ResNet50+Hilo attention	0.952245	0.907143	0.814285	1.0	0.897637	1.0
ResNet50+Hilo attention+ MSRNN	0.998776	0.971429	0.942857	1.0	0.970588	1.0
MobileNetV3	0.938571	0.821429	0.928571	0.714285	0.838709	0.764705
MobileNetV3+Hilo attention	0.957347	0.885714	0.914285	0.857142	0.888888	0.864864
MobileNetV3+Hilo Attention+MSRNN	0.994082	0.942857	0.942857	0.942857	0.942857	0.942857

basic models based on the model, as shown in Table 4. ResNet50 and MobileNetV3 are the basic feature extraction modules utilized, followed by the evaluation of the performance of the Hilo attention and MSRNN modules based on these two modules. According to the data in the table, it is concluded that the accuracy (ACC) of the Hilo attention module increases by 6.5%, the overall increase of Hilo attention + MSRNN is 12.9%, and the use of ResNet yields a 2.8% higher accuracy than that of MobileNet.

In Fig. 8, the model configuration of ResNet + Hilo attention + MSRNN exhibited the best patient-level performance, achieving an accuracy (ACC) of 0.9714, sensitivity (SEN) of 0.9429, specificity (SPE) of 1.0, area under the ROC curve (AUC) of 0.9988, and F1 score of 0.97. As depicted in Figure 9, all 70 patients with lung cancer were correctly predicted, while among the 70 normal subjects, four cases were incorrectly predicted as lung cancer patients.

V. CONCLUSION

The method proposed in this paper effectively characterizes the deep features of lung cancer lesions in CT images, enabling accurate differentiation between lung cancer patients and normal individuals in a weakly supervised manner. In this process, deep learning serves as a feature extractor, while Multiple Instance Learning (MIL) acts as a classifier, with the two approaches combined synergistically. By leveraging Hilo attention and MSRNN, the model learns 2D and 3D features from any number of CT images. Experimental results demonstrate that the method can aggregate both available and latent diagnostic features by exploiting the depth information of these features. With its practical application value established, it is anticipated that this method will exhibit excellent performance in other domains as well.

REFERENCES

[1] F. Manapov, C Eze, A Holzgreve. "PET/CT for Target Delineation of Lung Cancer Be-fore Radiation Therapy," Seminars in Nuclear Medicine, vol.52, pp.673-680, 2022.

[2] R. Manafi-Farid, E. Askari, I. Shiri. "[18F]FDG-PET/CT Radiomics and Artificial Intelligence in Lung Cancer: Technical Aspects and Potential Clinical Applications," Seminars in Nuclear Medicine, vol.52, pp.759-780, 2022.

[3] A. L. Potter, A. L. Rosenstein, M. V. Kiang. "Association of computed tomography screening with lung cancer stage shift and survival in the United States: quasi-experimental study," BMJ, vol.376, pp. e069008, 2022.

[4] B. M. Abunahel, B. Pontre, H. Kumar. "Pancreas image mining: a systematic review of radiomics," European radiology, vol.31, pp. 3447-3467, 2021.

[5] G. Chassagnon, C. Margerie-Mellon De, M. Vakalopoulou. "Artificial intelligence in lung cancer: current applications and perspectives," Japanese Journal of Radiology, vol.41, pp.235-244, 2023.

[6] K. He, X. Zhang, S. Ren. "Deep Residual Learning for Image Recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, vol.2016, pp.770-778, 2016.

[7] V. Ambrosini, S. Niccolini, P. Caroli. "PET/CT imaging in different types of lung cancer: An overview," European Journal of Radiology, vol.81, pp.988-1001, 2012.

[8] S. K. Thakur, D. P. Singh, J. Choudhary. "Lung cancer identification: a review on detection and classification," Cancer and Metastasis Reviews, vol.39, pp.989-998, 2020.

[9] W. Shen, M. Zhou, F. Yang. "Multi-scale Convolutional Neural Networks for Lung Nodule Classification," Information Processing in Medical Imaging: 24th International Conference, vol.2015, pp.588-599, 2015.

[10] Y. Xie, J. Zhang, Y. Xia. "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT," Information Fusion, vol.42, pp.102110, 2018.

[11] D. Ardila, A. P. Kiraly, S. Bharadwaj. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," Nature Medicine, vol.25, pp.954-961, 2019.

[12] A. Gopinath, P. Gowthaman, L. Gopal. "Enhanced Lung Cancer Classification and Prediction based on Hybrid Neural Network Approach," 2023 8th International Conference on Communication and Electronics Systems (ICCES), vol.2023, pp.933-938, 2023.

[13] H. Xiao, Q. Liu, L. Li. "MFMANet: Multi-feature Multi-attention Network for efficient subtype classification on non-small cell lung cancer CT images," Biomedical Signal Processing and Control, vol.84, pp.104768, 2023.

[14] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles," Artificial Intelligence, vol.89, pp.31-71, 1997.

[15] O. Maron, T. Lozano-Pérez. "A framework for multiple instance learning," Advances in Neural Information Processing Systems, vol.10, pp.570-576, 1998.

[16] M. Ilse, J. Tomczak, M. Welling. "Attention-based Deep Multiple Instance Learning, " Proceedings of Machine Learning Research, vol.80, pp.2127-2136, 2018.

[17] M. Tan, Q. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International conference on machine learning, vol.2019, pp.6105-6114, 2019.

[18] Y. Xu, H. Zhang. "Convergence of deep convolutional neural networks," Neural Networks, vol.153, pp.553-563, 2022.

[19] W. Alakwaa, M. Nassef, A. Badr. "Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)," International Journal of Advanced Computer Science and Applications, vol.8, pp.8, 2017.

[20] Z. Pan, J. Cai, B. Zhuang. "Fast vision transformers with hilo attention," Advances in Neural Information Processing Systems, vol.35, pp.14541-14554, 2022.

[21] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu. "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE transactions on medical imaging, vol.35, pp.129-1312, 2016.

[22] J. Chen, H. Zeng, C. Zhang. "Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics," Medical Physics, vol.49, pp.3134-3143, 2022.

[23] S. Tyagi, S. N. Talbar. "LCSCNet: A multi-level approach for lung cancer stage classification using 3D dense convolutional neural networks with concurrent squeeze-and-excitation module," Biomedical Signal Processing and Control, vol.80, pp.104391, 2023.

[24] S. Tomassini, N. Falcionelli, P. Sernani. "Cloud-YLung for Non-Small Cell Lung Cancer Histology Classification from 3D Computed

Tomography Whole-Lung Scans," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), vol.2022, pp.1556-1560, 2022.
[25] M. Ilse, J. M. Tomczak, M. Welling. "Deep multiple instance learning for digital histopathology," Academic Press, vol.2020, pp.521-546, 2020.
[26] X. Shi, F. Xing, Y. Xie. "Loss-Based Attention for Deep Multiple Instance Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol.34, pp.5742-5749, 2020.
[27] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles," Artificial Intelligence, vol.89, pp.31-71, 1997.
[28] S. Qi, C. Xu, C. Li. "DR-MIL: deep represented multiple instance learning distinguishes COVID-19 from community-acquired pneumonia in CT images," Computer Methods and Programs in Biomedicine, vol.211, pp.106406, 2021.
[29] N. Park, S. Kim. "How Do Vision Transformers Work?," arXiv, vol.2202, pp.06709, 2022.
[30] S. Merity, N. S. Keskar, R. Socher. "Regularizing and Optimizing LSTM Language Models," arXiv preprint arXiv, vol.1708, pp.02182, 2017.
[31] K. Clark, B. Vendt, K. Smith. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," Journal of Digital Imaging, vol.26, pp.1045-1057, 2013.