A graph neural network model for the diagnosis of lung adenocarcinoma based on multimodal features and an edge-generation network

Ruihao Li[1], Lingxiao Zhou[2], Yunpeng Wang[3], Fei Shan[4], Xinrong Chen[1], Lei Liu[1,5,6]

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China; [2]Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen, China; [3]Institutes of Biomedical Sciences, Fudan University, Shanghai, China; [4]Shanghai Public Health Clinical Center and Institutes of Biomedical Sciences, Fudan University, Shanghai, China; [5]Intelligent Medicine Institute, Fudan University, Shanghai, China; [6]Shanghai Institute of Stem Cell Research and Clinical Translation, Shanghai, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: All authors; (III) Provision of study materials or patients: F Shan; (IV) Collection and assembly of data: R Li; (V) Data analysis and interpretation: R Li; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Lei Liu, PhD. Academy for Engineering & Technology, Fudan University, Shanghai, China; Intelligent Medicine Institute, Fudan University, Shanghai, China; Shanghai Institute of Stem Cell Research and Clinical Translation, Fudan University, 138 Yixueyuan Rd., Shanghai 200032, China. Email: liulei_sibs@163.com; Xinrong Chen, PhD. Academy for Engineering & Technology, Fudan University, 220 Handan Rd., Shanghai 200433, China. Email: chenxinrong@fudan.edu.cn; Lingxiao Zhou, PhD. Institute of Microscale Optoelectronics, Shenzhen University, 3688 Nanhai Avenue, Shenzhen 518000, China. Email: lingxiaoz@szu.edu.cn.

Background: Lung cancer is a global disease with high lethality, with early screening being considerably helpful for improving the 5-year survival rate. Multimodality features in early screening imaging are an important part of the prediction for lung adenocarcinoma, and establishing a model for adenocarcinoma diagnosis based on multimodal features is an obvious clinical need. Through our practice and investigation, we found that graph neural networks (GNNs) are excellent platforms for multimodal feature fusion, and the data can be completed using the edge-generation network. Therefore, we propose a new lung adenocarcinoma multiclassification model based on multimodal features and an edge-generation network.

Methods: According to a ratio of 80% to 20%, respectively, the dataset of 338 cases was divided into the training set and the test set through 5-fold cross-validation, and the distribution of the 2 sets was the same. First, the regions of interest (ROIs) cropped from computed tomography (CT) images were separately fed into convolutional neural networks (CNNs) and radiomics processing platforms. The results of the 2 parts were then input into a graph embedding representation network to obtain the fused feature vectors. Subsequently, a graph database based on the clinical and semantic features was established, and the data were supplemented by an edge-generation network, with the fused feature vectors being used as the input of the nodes. This enabled us to clearly understand where the information transmission of the GNN takes place and improves the interpretability of the model. Finally, the nodes were classified using GNNs.

Results: On our dataset, the proposed method presented in this paper achieved superior results compared to traditional methods and showed some comparability with state-of-the-art methods for lung nodule classification. The results of our method are as follows: accuracy (ACC) =66.26% (±4.46%), area under the curve (AUC) =75.86% (±1.79%), F1-score =64.00% (±3.65%), and Matthews correlation coefficient (MCC) =48.40% (±5.07%). The model with the edge-generating network consistently outperformed the model without it in all aspects.

Conclusions: The experiments demonstrate that with appropriate data=construction methods GNNs

can outperform traditional image processing methods in the field of CT-based medical image classification. Additionally, our model has higher interpretability, as it employs subjective clinical and semantic features as the data construction approach. This will help doctors better leverage human–computer interactions.

Introduction

According to a 2020 report by the International Agency for Research on Cancer, lung cancer is the leading cause of cancer-related death among all gender groups (18%) and the second leading cause of new cancer cases (11.4%) (1). Implementing an early screening program to diagnose patients is one of the major steps to reducing lung cancer-related death and improving survival (2). Both the US National Lung Screening Trial (NLST) and Dutch-Belgian Lung Cancer Screening Trial (NELSON) concluded that lung cancer mortality could be significantly reduced by early screening and self-management through low-dose computed tomography (LDCT) in high-risk lung cancer populations (3,4).

As the most common histological subtype of the lung cancer, lung adenocarcinoma accounts for about half of lung cancers. According to the international gold standard, adenocarcinoma is divided into 4 categories: atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC) (5,6). Several clinical studies (7,8) have demonstrated that nodules with a diameter of less than 5 mm are more likely to be benign, with a malignant risk typically less than or equal to 1%. Conversely, according to several investigations (9,10), the majority (80%) of cancerous nodules have a size greater than 8 mm, and there is evidence to suggest that nodules smaller than 10 mm may also be malignant IAC nodules with a chance of spreading (11,12). We believe that nodules in the 5 to 10 mm range are clinically challenging to discern and include various subtypes of pulmonary adenocarcinoma nodules. Recently, it has been shown that CT-based handcrafted and deep radiomics has been able to determine the invasiveness of lung adenocarcinoma and that a combination of other variables (such as clinical, semantic, and pathological features) can improve the accuracy of the final pathology (13). Therefore,

developing a multimodal feature CT diagnosis system for nodules in the 5 to 10 mm range is necessary. Typically, 3 categories of features are used in diagnosis: semantic features, radiomics features, and deep features.

The radiologist will usually describe and analyze the lesion by qualitative or quantitative semantic features, which usually include shape, location, lobulation, size, volume, etc. However, these methods usually require a highly skilled clinician and rely on subjective appraisal; thus, a series of terms and gold standards need to be specified. Based on evidence indicating that semantic features of CT images have prognostic value, the Lung Imaging Reporting and Data System (Lung-RADS) has been developed to improve the interpretability of lung cancer screening CT images and facilitate the prognostic management of cases (14). Undoubtedly, semantic features are often limited by the subjectivity of evaluation, making model consistency difficult to achieve (15).

As image processing technology matured, radiomics was first proposed by the Dutch researcher Lambin *et al.* in 2012 (16). Although it can only extract low-level features, radiomics is nonetheless an automatic, high-throughput feature extraction method for transforming images into minable feature data, providing far more consistency than semantic features. Radiomics features are outputs of the values of image pixels in the region of interest (ROI) that are input into mathematical formulae. In radiomics, traditional image features, such as shape, grayscale, and texture, are extracted, after which pattern recognition models are applied for classification and prediction (17). Currently, several standardized software packages for extracting radiomics features have been developed, with the Python-implemented PyRadiomics (18) being one of the most well-known and widely used in nodule feature extraction (19).

Research in deep learning has progressed slowly over

time, and it is only in the past decade that significant breakthroughs have been made, with advancement from LeNet (20) to AlexNet (21) and then to a series of deep learning algorithms. The extraction of deep features usually requires convolution, pooling, activation, and full connection layers. Features in the shallow layer are similar to those of radiomics features, but with the deepening of the layers, the features become increasingly abstract and more laden with high-level information; however, due to this higher dimensionality, these features are more difficult to interpret. This makes the magnitude of the required data much larger than that required by radiomics models. A convolutional neural network (CNN) is thus often used for image analysis and deep feature extraction of Euclidian data.

The classification of lung adenocarcinoma nodule subtypes on CT is mostly based on CNN or machine learning methods. Wang *et al.* [2021] (22) proposed a method in which a mask segmentation model built with a 3D U-Net and a classification model of 3D convolution were jointly trained to complete the classification of subtypes. Yu *et al.* [2021] (23) used a 3D multimask network to determine the invasiveness of ground-glass nodules (GGNs). Ashraf *et al.* [2022] (24) used a 3D multiscale CNN composed of 3 cubic subvolumes from computed tomography images to predict whether a lung nodule was benign, adenocarcinoma, or a preinvasive subtype.

In the field of deep radiomics, Paul *et al.* [2018] (25) fused deep features and radiomics features to classify benign and malignant pulmonary nodules with a smaller set of training data than that of a radiomics-only method, achieving superior performance. Xia *et al.* [2020] (26) combined a deep feature–based model score obtained by transfer learning with a radiomics feature-based model to classify IAC and non-IAC. Wang *et al.* [2020] (27) used multitask and 3D-convolution models of deep radiomics to distinguish IAC from non-IAC, MIA from AIS, and MIA from IAC. Their results showed that the deep-radiomics models perform better than do deep-learning–based models and classification and segmentation models. Wang *et al.* [2021] (28) used the combination of depth, radiomics, and natural language–processed (NLP) pathological features for classifications of 2, 3, 6, and 8 categories.

Since the datasets of medical images are usually small in magnitude, we propose a series of multimodal feature methods for deep radiomics to better fit deep learning and supplement the shallow feature space. In addition, as a solution to the issues of poor interpretability in deep-radiomics models and poor consistency in semantic models,

we then fuse semantic features into our feature sets, forming a combination of 3 features (semantic, deep, and radiomics features). After consideration, we believe a graph neural network (GNN) is more suitable for merging these 3 features.

GNN is a deep-learning-based method that runs on graph domains. GNNs have recently become a widely used graphical analysis method due to their excellent performance (29). A graph is a data structure that models a set of objects (nodes) and their relationships (edges). As a unique form of machine learning with non-Euclidean data structures, graph analysis focuses on tasks such as node classification, link prediction, and clustering. GNNs can be divided into 2 types according to the definition method of the graph convolution operator. The first is the spectral domain-based definition of the graph convolution operator, with the representative method for this type being a graph convolutional network (GCN) (30). The second method is the definition of a graph convolution operator based on the spatial domain. The spatial method aims to aggregate each central node and neighboring nodes by defining aggregation functions from the node domain. Typical methods of this type include the graph attention network (GAT) (31) and GraphSAGE (32). Among the GNNs used for medical images, a relatively classic work is that of Parisot *et al.* [2018] (33), in which the authors input the extracted image features into a graph composed of phenotypic data for edge weights; the nodes were then classified by GCN for the semisupervised learning of Alzheimer disease. In the work of Hao *et al.* [2022] (34), a Bayesian CNN was used to extract image features; these were then combined with uncertainty to complete graph characterization for the constructed CT maps and ultimately classify pulmonary effusion.

The unique advantage of graph data is that the data contain critical relational information, which Euclidean data typically lack. By utilizing graph data, we can accomplish 3D feature extraction in a 2D manner while also facilitating the integration of multimodal features. Therefore, in the process of applying GNNs to Euclidean data, this is crucial for the establishment of data connection. The construction of the graph data structure of medical images is always challenging, which is one of the reasons why GCNs are rarely used in medical research. Meanwhile, in the process of the subjective construction of a graph data structure, there may be nodes without out-degree or in-degree, which may hinder the overall classification and prediction performance. This is another problem that we encountered

Table 1 Training and test set ratio

Categories	Training set (80%)	Test set (20%)	Total
MIA	154	39	193
IAC	69	17	86
AIS	47	12	59
Total	270	68	338

MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; AIS, adenocarcinoma in situ.

in building the data.

The following describes our solution to the above problems: deep features, radiomics features, and semantic features are combined to classify patients into 3 categories of lung adenocarcinoma. The semantic features provide interpretability, the deep features extract the high-level features of CT, and the radiomics features extract the low-level features of CT. Additionally, we combine the deep features and radiomics features to ensure the consistency of the features and use an edge-generation network to fix the graph data structure.

The novel contributions of this work are as follows:

(I) We demonstrate the superiority of multimodal fusion features in the subtyping of lung adenocarcinoma and have introduced a graph-based approach for multimodal feature fusion of CT images, which involves the integration of deep, radiomics, and semantic features.

(II) In the process of constructing graph data, subjective semantic features were used to determine the connectivity between nodes, resulting in a more transparent information propagation path between nodes and improved interpretability for clinical applications.

(III) We trained an edge-generation network to predict the edge connectivity between nodes without outgoing or incoming edges, which completes the information of graph data and improves the performance and stability of the model. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-2/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the ethics committee of the Shanghai Public Health Clinical Center. Informed consent was obtained from all the patients.

The following section describes the screening of the dataset and the proposed framework.

Dataset

The data were obtained from 668 patients attending Zhongshan Hospital and Shanghai Public Health Clinical Center between September 2015 to July 2019. The inclusion criteria for data were as follows: (I) patients underwent routine chest CT at 2 hospitals within 1 month before resection, (II) patients were examined using CT with a slice thickness less than or equal to 1 mm, (III) patients had no breath motion artifact, (IV) nodules were between 5 and 10 mm, (V) nodules were ground-glass type, and (VI) the relation between the mask and raw CT was clear and true.
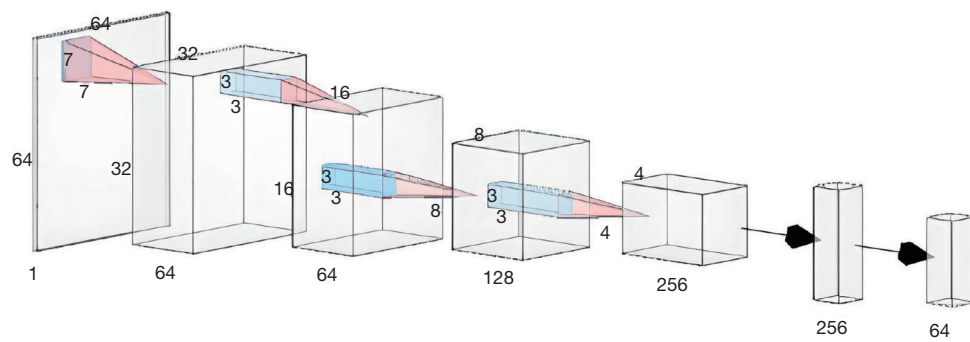
After data screening, we included a total of 338 nodules from 307 patients (some from the same patient but of different subtypes) 5–10 mm in size, consisting of 86 cases of IAC, 193 cases of MIA, and 59 cases of AIS. According to Son *et al.* [2016] (6), the persistent presence of GGNs in CT images is usually a sign of the presence of lung adenocarcinoma or its precursors. Therefore, all samples included were GGNs, with 112 mixed GGNs (mGGNs) and 226 pure GGNs (pGGNs). As we employed 5-fold cross-validation and the total number of data could not be divided by 5, there is a slight difference in the number of samples between the training and testing sets in each fold. The ratio of the training set to the testing set is 4:1, and they have the same distribution, as shown in *Table 1*.

Measurements were taken by 2 radiologists with more than 5 years of experience in the field of chest radiology. The 2 radiologists measured each image feature separately, and a third radiologist with more than 20 years of experience in the field of chest radiology reassessed any discrepant cases. The masks were reviewed by a radiologist with more than 6 years of experience in the field of chest radiology and by a radiologist with more than 20 years of experience in the field of chest radiology.
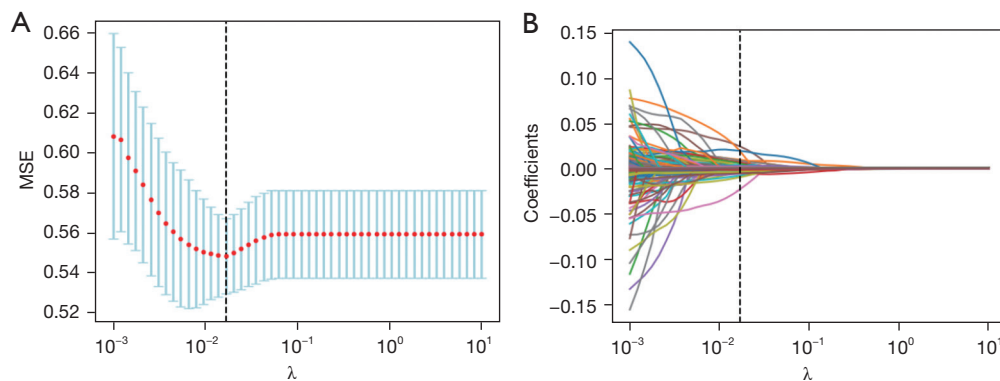
Method overview

The framework of the proposed approach consists of 3 modules: a feature extraction module, feature fusion module, and classification module. The overall framework is shown in *Figure 1*. Our work is a supervised classification

Figure 1 The overall framework of our work, including the feature extraction module, feature fusion module, and classification module. CT, computed tomography; ROI, region of interest; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; AIS, adenocarcinoma in situ; GCN, graph convolutional network; LASSO, least absolute shrinkage and selection operator.

model aimed at integrating multimodal features of pulmonary nodule CT images for classification. First, we employed ResNet (35) as the backbone network to extract deep features and used the PyRadiomics library, an open platform built in Python (Python Software Foundation), to extract radiomics features, while clinical and semantic features were annotated by experienced radiologists in advance. Subsequently, we employed a graph data structure to simulate the slice scan structure of CT images and fused all features. Moreover, we constructed a new graph dataset with semantic features with an edge-generation network, where each node represented a case of pulmonary nodules. Finally, traditional GNNs were used for node classification.

Feature extraction module

Deep feature extraction

In the deep feature extraction part, after consideration, we decided to use ResNet34 as the feature extraction network to extract the deep features of each slice of the CT images for all nodules. However, due to the issue of image scale, we removed the last module of ResNet34 to prevent the convolution kernel size from exceeding the image size. Consequently, the modified ResNet was named *ResNet28*. The parameters of the deep feature extraction network were obtained by using ResNet28 as the deep learning model. The model was trained for 200 epochs on the augmented ROI image data, and the parameters from the epoch with

5338

Li et al. A lung nodule diagnosis model based on GNNs

Figure 2 Deep feature extraction with ResNet28.

Figure 3 Radiomics selection using the LASSO regression model. (A) The best λ value was selected according to the minimum MSE. (B) Through 10-fold cross validation, 38 features were finally selected from 716 characteristic coefficient curves using λ values. LASSO, least absolute shrinkage and selection operator; MSE, mean square error.

the best results were selected. We cropped ROIs 64×64 in size from the center of the nodules, and these cropped ROIs were used as the input for the feature extraction part. A series of convolution layers, batch normalization layers, and rectified linear unit (ReLU) layers was used, and after flattening was performed by the fully connected layer, the extracted result was a 64-dimensional tensor. The structure of ResNet28 is shown in *Figure 2*.

Radiomics feature extraction

Traditional radiomics features were extracted for each slice (dimension of pixel is 512×512×1) in CT through the PyRadiomics library in Python and the least absolute shrinkage and selection operator (LASSO) regression method in the scikit-learn library (36,37). Finally, the extracted result for each slice was a 38-dimensional tensor, in which LASSO determined the value of λ according to the minimum mean squared error. The result of LASSO is shown in *Figure 3*. After the λ was determined, 38 stable radiomics features from 716 features were finally saved for classification.

Feature fusion

Fused feature representation

After a concatenating operation is completed, deep features and radiomics features are input into the graph with a similar fixed structure with self-connection for feature fusion, as shown in *Figure 4*. This structure was first proposed by Hao *et al.* [2022] (34), and in our work, all the information in every slice is aggregated at the central node, which means that the information in the central node can be viewed as all the information extracted from the nodule's CT image. Although all the features were extracted in a 2D manner, we could also obtain pseudo-3D results using this graph structure. Moreover, this structure effectively incorporates spatial information and transfers information between adjacent and subadjacent slice nodes. Due to the presence of self-connections, each node can also effectively emphasize its own information. All the information in each slice is aggregated in the central node, which can be regarded
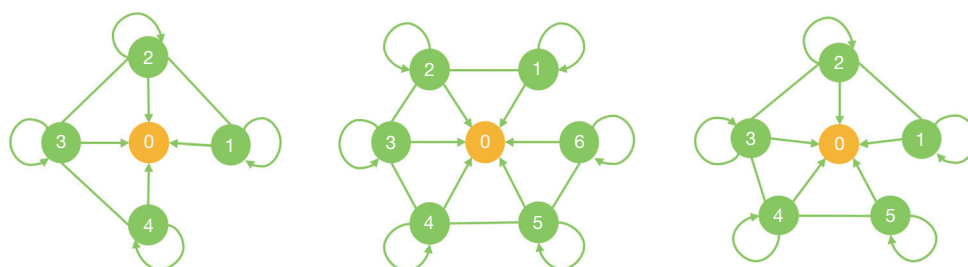
Figure 4 In the figure, each node represents a CT slice. The structure of the directed graph is adopted. Each node only has a directed edge with its adjacent node and central node, and each node has self-connection. At the very center is the fictitious spatial structure node, initialized with tensors of the corresponding dimension that are all zeros. CT, computed tomography.

as the summary of all the information extracted from the CT images of the nodule. Another advantage of the construction is that it is not necessary for all data to be of the same size. The feature fusion can be completed by only intercepting the relevant slice information of each nodule. The flexibility is improved compared to methods that must extract a fixed size, such as 3DCNN, and further saves memory for computing.

Another reason for using fusion is that in the subsequent GNN, training is carried out on a per-nodule basis, and a same-dimensional feature that can represent each nodule is needed rather than the deep features or radiomics features extracted from each slice. The fused feature is only the feature vector in the central node. The node representations of the graph are generated using the information aggregation structure of the graph isomorphism network (GIN) (38). The GIN is a spatial-based convolutional GNN inspired by the Weisfeiler-Lehman (WL) test and is essentially used for distinguishing isomorphic graphs. GIN is employed in this method to make the nodules with similar sizes more similar in terms of information aggregation and representation, which naturally incorporates the semantic feature of nodule size.

In our proposed method, since the shape of the graph is predetermined and only the feature information of the nodes needs to be aggregated and since the weights of each node in the CT are not the same, the pooling part of the graph readout is abandoned, and only the aggregated result in the central node is used as the embedding vector of the graph representation. Furthermore, considering the interpretability (in general, it is believed that a slice is only related to, at most, 2 adjacent slices above and below it), we only aggregate the neighbor node features within 2 steps for each node and only take the aggregated result of the last graph convolution layer. The feature aggregation function

of GIN is as follows:

$$h_v^{(k)} = MLP^{(k)}\left(\left(1+\varepsilon^{(k)}\right)h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)}\right) \quad [1]$$

where ε is a learnable parameter, and $h_v^{(k)}$ is the aggregate result of layer k about node v.

Graph construction and node classification
Construction of graph data

The fused feature is input into the graph database established by 9 class semantic features, which include type, size, spiculation, lobulation, vacuole sign, air bronchogram, vessel (normal or abnormal), tumor–lung interface (clear or not), and pleural indentation. Node classification can be performed using GAT or GCN. According to the study conducted by Hu *et al.* [2021] (39), the risk of IAC increases in nodules over 8mm, and the same increased risk is applicable to nodules with spiculation, lobulation, vacuole sign, vessel abnormalities, or pleural indentations. According to our experience and the advice of the radiologist, we added 2 signs of air bronchogram and an obvious tumor–lung interface. Ultimately, the similarity judgment was mainly divided into 3 aspects: (I) similarity is plus 1 if 2 nodules have the same type; (II) the threshold for similarity determination is a size of 8 mm; (III) if 3 or more signs of the remaining 7 signs are the same between 2 nodules, every sign makes the similarity plus 1. Subsequently, edges in the graph are connected based on the magnitude of their similarity, with a similarity threshold of 5 being used in the model. As the connections between nodes were made using subjective methods, we can provide explanations for the propagation of information in the graph. Therefore, the graph constructed in this work is an interpretable directed graph with 338 nodes and 21,258 edges. It is also possible to customize the edge connection in a way that is considered more interpretable,

5340

Li et al. A lung nodule diagnosis model based on GNNs
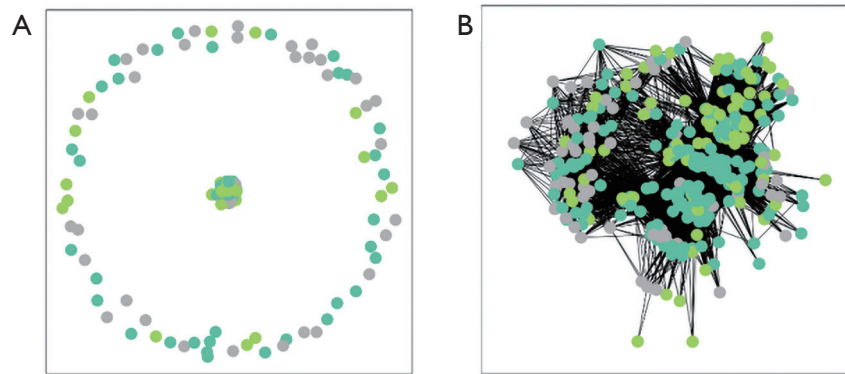
Figure 5 Visible graph structure with and without an edge-generation network. (A) Visible graph structure before edge generation and (B) visible graph structure after edge generation.

demonstrating the flexibility of our method.

Edge-generation network

After the graph data structure is constructed, due to its high subjectivity, it must be accompanied by certain limitations; otherwise, as shown in *Figure 5A*, the graph would not be fully connected.

We refer to the dense middle part the of the graph network, in which the nodes are all connected, as the backbone. We can find that in the graph data structure constructed purely subjectively, there are many nodes left unconnected by edges with the backbone graph network, which means that their self-features cannot be updated in the process of feature aggregation. Therefore, to reduce the number of unconnected nodes and maximize the effect of feature aggregation, we propose an edge-generation network that is achieved through the following steps.

(I) Find all nodes that are not connected to the backbone graph;

(II) Take all nodes in the backbone as positive examples of the training set and extract the virtual edges as negative examples for training to learn the mode of the subjective connection. The loss function of training is as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log(x_n) + (1-y_n)\log(1-x_n)\right] \quad [2]$$

where N is the number of samples, y_n is the symbolic function, and x_n is the prediction probability.

(III) Encode the overall image composed of all nodes. The encoder uses 2 layers of GCN convolution, where the hidden layer has a dimension of 128.

(IV) Decode the output of the encoder.

(V) Predict the probability of edge existence between nodes using the sigmoid, with the α edges with being highest probability being selected for completion of the graph connection. A new graph, as shown in *Figure 5B*, can be generated in this way.

Experiments

In this section, our intra-, inter, and ablation experiments are reported and discussed. We fixed the random seed of all experiments to 42 and all the learning rates of models to 0.001. The dataset was split in the manner shown in *Table 1* for all experiments except for deep-feature model with data augmentation in inter-experiments. To address the issue of imbalanced data, we introduced class weights into the loss function for each category based on its frequency in the dataset. Therefore, the loss function is as follows:

$$L_{CE} = -\frac{1}{M}\sum_{i}\sum_{c=1}^{M} y_{ic} \log(p_{ic}) w_c \quad [3]$$

where M is the number of categories, which in this experiment was 3; y_{ic} is a symbolic function that takes 1 while the actual class of sample i equals c, and 0 otherwise; p_{ic} is the prediction probability that sample i belongs to class c; and w_c is the weight of category c.

Metrics

We used accuracy, F1-score, Matthews correlation coefficient (MCC), and area under the curve (AUC) as the metrics in our experiments. All the metrics were averaged after 5-fold cross-validation. In each fold, the macroaverage

was used for calculation, the formulae for which are as follows:

$$ACC = \frac{TP + TN}{N} \qquad [4]$$

$$F1 = \frac{2 \times P \times R}{P + R} \qquad [5]$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad [6]$$

where P is precision, R is recall, TP is true positive, TN is true negative, FP is false positive, FN is false negative, and N is the total number of samples.

Interexperiments

This section describes the models used in the comparison with our proposed model and include the deep-feature model, radiomics feature model, and semantic models.

Deep-feature model

The first baseline model is ResNet28, as presented in *Figure 2*, to which we added a fully connected layer at the end. The second baseline was Local-Global Net (40), and the third baseline was 3D neural architecture search (3D-NAS) (41), both of which are classic algorithms for lung nodule classification. The optimizer for all these models is Adam, and there are 200 epochs in total in each fold. The first 2 models require 2D input, while the last model requires 3D input. For 2D input, an ROI of size 64×64 was cropped from the center of each nodule's maximal axial slice. For 3D input, a cubic ROI of size 32×32×32 was cropped, which included the nodule in the center. Data augmentation was performed by applying flipping, rotation (90°, 180°, and 270°), and Gaussian blurring to each ROI. As a result, the dataset used in this study was 6 times larger than that used in other models.

Radiomics feature model and semantic feature models

We employed support vector machine (SVM), random forest (RF), and adaptive boosting (AdaBoost) algorithms for the classification of radiomics features and semantic features. For SVM, we used a linear kernel with a penalty coefficient of 1.0. The RF algorithm consisted of 50 decision trees as base classifiers. The base classifier for AdaBoost was decision tree, and the number of the base classifiers was set to 50.

Intraexperiments

All the models described in this section share the same structure, with differences only in the classification network or hyperparameters.

GNN model

For the graph node classification task, we tried 2 well-regarded GNNs model, namely GCN and GAT (heads =3). The depth of the network is 2 layers, and the dimension of the hidden layer is 64. Each model has a dropout of 0.5.

Hyperparameter α

In the experiment, we constantly adjusted the hyperparameter α to find the appropriate graph data completion method. The function of the hyperparameter is to determine the number of edges with the highest probability of connection between the unconnected nodes and other nodes. The probability of an edge between nodes is obtained by an edge generator.

Ablation experiments

Deep-feature model and deep-radiomics feature model

In the ablation experiments, during the feature fusion part, the radiomics features were removed, and only the deep features were used for fusion, which were subsequently sent to the next stage in order to prove the performance superiority of multimodal features over single-modal features.

Original model and edge generation model

In this experiment, we derived the detailed F1-scores for different categories in the multiclass task of GCN and GAT models with and without the edge generator. The purpose of this experiment was to demonstrate that the edge-generation network could enhance the overall performance of the model.

Results

The results of all models used in this study on our dataset can be found in *Table 2*, and the *t*-test result between models can be found in *Figure 6*. Without data augmentation, our model achieved the best performance among all models. The results of our method are as follows: ACC =66.26% (±4.46%), AUC =75.86% (±1.79%), F1-score =64.00% (±3.65%), and MCC =48.40% (±5.07%). When data augmentation was applied to the deep feature model, our model still outperformed the Local-Global Net in all performance metrics. The Local-Global Net is a model that performs well in pulmonary nodule classification by using different modules to extract local and global features. In comparison with ResNet28, our model only had a lower ACC, while the other performance metrics were slightly

Table 2 The overall comparison of classification results

Metrics	Fused feature		Deep feature			Deep feature with data augmentation			Radiomics feature			Semantic feature		
	Ours(O) GAT	Ours(G) GAT	3D-NAS	Local-Global Net	ResNet 28	3D-NAS	Local-Global Net	ResNet 28	SVM	RF	Ada-Boost	SVM	RF	Ada-Boost
ACC	65.99±4.23	66.26±4.46*	48.93±9.00	48.59±8.18	49.46±9.50	70.32±5.73#	58.43±2.41	67.31±1.35	57.70±1.41	57.66±3.20	54.09±5.02	59.16±1.88	52.07±2.99	56.21±9.52
AUC	71.34±3.15	75.86±1.79*	62.34±1.27	54.11±3.68	56.77±2.25	82.96±4.17#	71.20±2.40	70.65±0.88	60.30±4.85	66.45±3.07	61.70±4.48	73.40±4.51	71.41±2.31	72.82±5.73
F1	61.13±5.35	64.00±3.65*	40.36±5.00	34.17±5.60	35.87±3.39	67.31±4.57#	55.16±2.28	58.06±2.16	28.27±3.30	40.16±3.13	45.85±4.69	49.04±2.06	45.03±3.08	51.27±10.55
MCC	43.62±8.09	48.40±5.07*	22.08±2.79	14.88±5.86	17.68±6.92	50.42±5.98#	33.30±4.41	40.67±1.87	8.15±10.22	16.42±5.95	17.14±7.37	23.18±2.80	15.16±4.75	24.76±16.63

The results are the mean ± standard deviation of 5-fold validation. *, the best result without data augmentation; #, the best result after data augmentation. ACC, accuracy; F1, F1-score; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; GAT, graph attention network; NAS, neural architecture search; SVM, support vector machine; RF, random forest.

higher than those of ResNet28. However, our model was less stable than was ResNet28, indicating that our model and the ResNet28 model with data augmentation have highly comparable performance. The 3D-NAS, which is a low-computational complexity model based on 3D convolution, slightly outperformed our model in all the performance metrics. However, considering that our dataset was only one-sixth the size of the augmented dataset and that 3D-NAS is a state-of-the-art model that ranks highly in public datasets, we feel that this result is acceptable.

Table 3 presents the results of different GNNs as classifiers, along with the F1 scores of the 3 subtypes in each model. It can be observed that GAT outperforms GCN in overall performance on our dataset. Models with edge-generation networks perform better than do their counterparts without edge generators, particularly in the IAC and AIS minority classes, where the stability of models with edge a generator appears to be greatly improved.

Figure 7 presents the results of 4 metric indicators as a function of the hyperparameter α. The hyperparameter α regulates the number of edges generated by the edge-generating network. We only considered the range from 1 to 20 since nodes in the graph network become too significant and can distort the original judgments when the edge generation count exceeds 20. It can be observed that the optimal performance of our model could be obtained with α lying between 6 and 10. All GAT results presented in this paper were derived with α equal to 10.

Table 4 compares the results of models using multimodal feature fusion with those using only unimodal feature fusion. The results indicate that the models using multimodal feature fusion significantly outperformed those using only unimodal feature fusion.

Discussion

Traditional machine learning methods for lung nodule classification usually use radiomics and semantic features as inputs and employ machine learning classifiers for classification. However, such an approach can lead to a lack of deep features, which can affect the final results. Deep learning methods use deep models to extract image features and achieve better results but require large amounts of data and are difficult to interpret. Our model integrates multimodal features and constructs a graph data structure to fuse spatial information and semantic features, making the information propagation path clearer. When both the low-level features contained in the radiomics features and
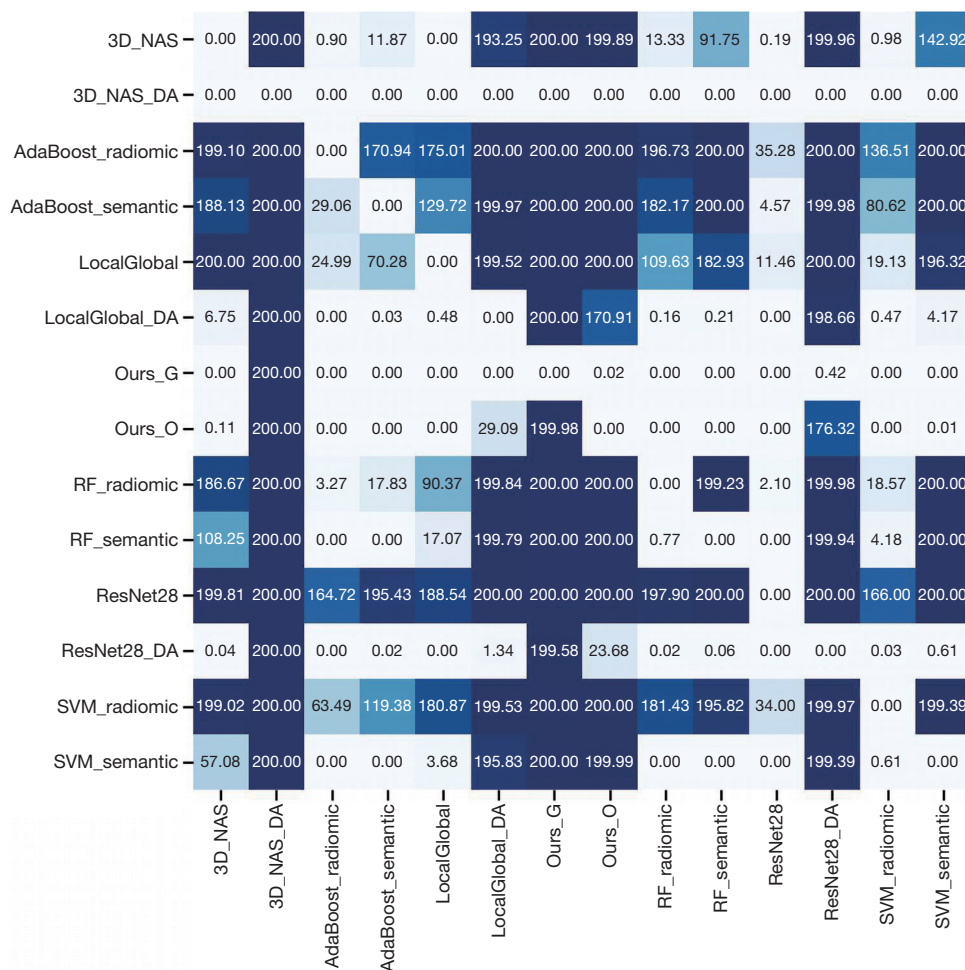
Figure 6 A heatmap of the *t*-test results between models. GAT, graph attention network; NAS, neural architecture search; SVM, support vector machine; RF, random forest; DA, data augmentation.

Table 3 Results of the GAT and GCN models

Metrics	GAT		GCN	
	Original	Edge generation	Original	Edge generation
ACC	65.99±4.23	66.26±4.46*	65.42±2.91	64.82±6.38
AUC	71.34±3.15	75.86±1.79*	71.50±3.00	73.68±6.58
F1	61.13±5.35	64.00±3.65*	57.60±5.98	60.84±6.62
MCC	43.62±8.09	48.40±5.07*	42.34±7.33	44.04±8.59
F1, MIA	72.67±5.78	72.71±5.61*	73.64±4.44	74.26±5.03
F1, IAC	55.30±18.72	61.00±3.20*	50.44±26.25	57.58±8.19
F1, AIS	55.43±8.32	58.28±3.01*	48.71±13.37	50.69±10.05

The results are the mean ± standard deviation of 5-fold validation. *, represents the best result without data augmentation. ACC, accuracy; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; F1, F1-score; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; AIS, adenocarcinoma in situ; GAT, graph attention network; GCN, graph convolutional network.

5344

Li et al. A lung nodule diagnosis model based on GNNs
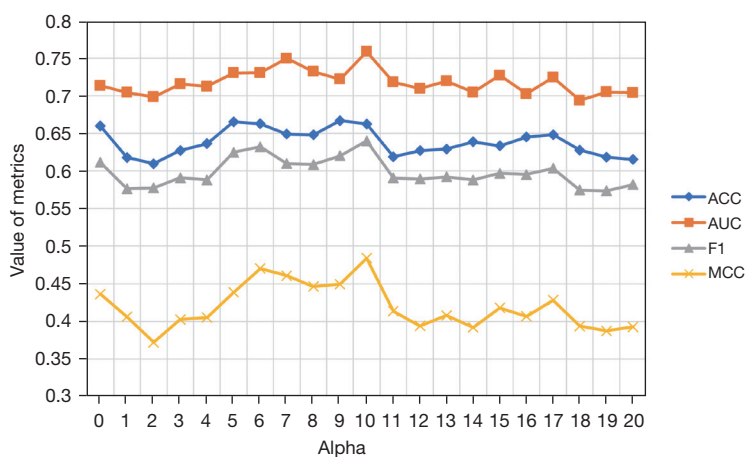
Figure 7 Relation between the hyperparameter value and the experimental metrics. ACC, accuracy; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; F1, F1-score.

Table 4 Results of the models with deep-radiomics features and those with deep features only

| Metrics	Deep-radiomics feature model	Deep-feature model
ACC	66.26±4.46	61.86±4.64
AUC	75.86±1.79	73.12±4.66
F1	64.00±3.65	57.9±3.89
MCC	48.40±5.07	40.60±4.50

The results are the mean ± standard deviation of 5-fold validation. ACC, accuracy; F1, F1-score; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve.

the high-level features extracted from the deep features are expressed, the overall features of CT images can be more completely presented and received by the classification network. Our model is not only richer in feature dimensions but also better in interpretability compared to models employing only deep features. Therefore, as shown in *Table 2*, our model demonstrated a superior performance to those of traditional machine learning algorithms, including SVM, RF, and AdaBoost. Compared with deep learning models, ours has advantages when the data volume is small. However, due to the demand of features being too high and the difficulty in generating semantic features, the data required for our work cannot be expanded through data augmentation. Therefore, our model is slightly inferior to the self-organizing tree algorithm (SOTA) method 3D-NAS, in which data augmentation is used. However, considering the difference in data volume, we believe that

this difference is reasonable, and we will seek to improve our model in future work.

To address the issue of data imbalance, we increased the weight of each sample in the model's loss function according to its proportion in the dataset. This led to improved performance of GAT over GCN in the selection of a node classifier. This is because GAT is an attention-based network that assigns different attention coefficients to each vertex. These coefficients are used to compute node representations, which in turn are used to compute the loss function and guide the model training process. When we added the weight to the loss function, the corresponding attention coefficients became more precise during the backpropagation, resulting in a more accurate model.

The use of an edge-generation network for information completion on semantic feature–constructed graphs is a highlight of this work. *Table 3* shows the comparative results between the original model and the model with the added edge-generation network, which validates the effectiveness of the edge-generation network. Compared with the original model, the model with the edge-generation network shows improvements in all performance metrics. Moreover, significant improvement in the stability of the minority class samples can also be observed. This is due to the existence of a large number of nodes without edges in the original graph data, and these nodes' information is usually not propagated to other nodes in GNNs because of the lack of relationships with other nodes. Thus, there is information loss during the information propagation process. The classification prediction can only be based on the node embedding, which is closer to random
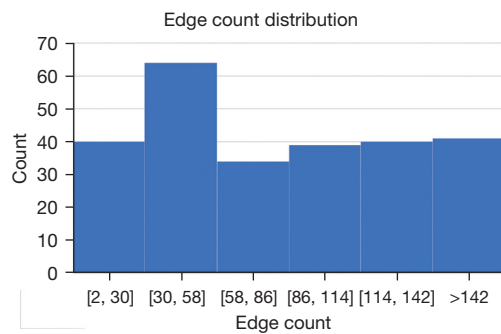
Figure 8 Distribution of edges of the original graph.

guessing, leading to high instability in the prediction results across different data distributions. However, when the edge-generation network learns and predicts the edge connections, the information can flow among these originally discrete nodes, which improves the stability. In the edge-generation network, the hyperparameter α is used to connect the originally discrete nodes to α adjacent nodes for feature information propagation. A too-large value of α may cause less important nodes to become central nodes in the information propagation, so the value of this hyperparameter needs to be manually set. In our original graph data, there are 258 nodes with edges and a total of 21,258 directed edges. The edge count distribution, shown in *Figure 8*, is close to a random distribution. Considering that these initially unconnected nodes are less important than are those with edges, we expect to afford them relatively fewer undirected edges. Therefore, we believe it is reasonable to set up to 20 adjacent edges for nodes without edges. The experimental results show that connecting each node with 6–10 edges is the optimal range.

In terms of the number of the model parameters, besides the 20 million parameters in the ResNet28 deep model used for extracting deep features, the proposed model only uses slightly over 40,000 parameters in the edge-generation network and 20,000 parameters in the GAT used as the classifier. This means that by adding the proposed subsequent structures and fusing multiple features on top of the deep network, better performance can be achieved in the multiclassification task of pulmonary nodules with almost no additional burden.

Although our model has many advantages, there are still limitations and room for improvement. First, the model uses multimodal features, including semantic features, making it difficult to perform data augmentation. Therefore, our model requires the collection of a large amount of medical data for

data augmentation. We will consider generating imaging features for pulmonary images in future work to support data augmentation for the model. Second, due to subjective factors, although the edge-generation network is used, the construction of the graph still becomes unstable depending on the different connection patterns, resulting in poor performance in domain adaptation. Therefore, our method still needs to be improved in terms of robustness. One direction for consideration is the optimizing pf the first-order and second-order similarity losses for graphs in different domains. Finally, our work in clinical automation relies on the accuracy of the multitask detection of pulmonary nodule signs, which is also an area we will investigate further.

Conclusions

For lung cancer, the detection and screening of early lung adenocarcinoma subtypes are critical. Compared to using a single type of feature, leveraging the fusion of multiple deep and shallow features with semantic features of CT confers substantially more benefit in terms of the comprehensiveness of information. Owing to the growing sophistication of sign detection technology, there are many upstream benefits to our work.

Compared with traditional CNN and machine learning methods, our method has better performance and a relatively smaller number of parameters. The graph network is undoubtedly a superior choice for the relationship processing between CT slices and the connection between similar cases due to its implicit spatial relationship attribute. This enables our work to introduce spatial relationships while eliminating the disadvantage of large memory requirements for 3D convolution. In addition, due to the empirical subjective opinions introduced by the construction of subjective graph data and the supplement of the graph structure by the edge-generation network, the data construction method, similarly to hyperparameters, provides clinicians a more intuitive interpretability and better interactivity with computers.

To the best of our knowledge, this is the first attempt to combine GNN with multimodal features for the multiclassification of lung adenocarcinoma nodule subtypes. We hope that our work leads to further developments in this field.

Acknowledgments

First-Class University Construction and Development of High-Level Local Universities, the Academy for Engineering and Technology of Fudan University, the Intelligent Medicine Emerging Interdisciplinary Cultivation Project and Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai for financial assistance and technical advice.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-2/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-2/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the ethics committee of the Shanghai Public Health Clinical Center. Informed consent was obtained from all the patients.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.

2. Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. Lancet 2021;398:535-54.

3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.

4. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med 2020;382:503-13.

5. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011;6:244-85.

6. Son JY, Lee HY, Kim JH, Han J, Jeong JY, Lee KS, Kwon OJ, Shim YM. Quantitative CT analysis of pulmonary ground-glass opacity nodules for distinguishing invasive adenocarcinoma from non-invasive or minimally invasive adenocarcinoma: the added value of using iodine mapping. Eur Radiol 2016;26:43-54.

7. Horeweg N, van Rosmalen J, Heuvelmans MA, van der Aalst CM, Vliegenthart R, Scholten ET, ten Haaf K, Nackaerts K, Lammers JW, Weenink C, Groen HJ, van Ooijen P, de Jong PA, de Bock GH, Mali W, de Koning HJ, Oudkerk M. Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening. Lancet Oncol 2014;15:1332-41.

8. Larici AR, Farchione A, Franchi P, Ciliberto M, Cicchetti G, Calandriello L, Del Ciello A, Bonomo L. Lung nodules: size still matters. Eur Respir Rev 2017.

9. Henschke CI, Yankelevitz DF, Mirtcheva R, McGuinness G, McCauley D, Miettinen OS; . CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. AJR Am J Roentgenol 2002;178:1053-7.

10. Swensen SJ, Jett JR, Hartman TE, Midthun DE, Sloan JA, Sykes AM, Aughenbaugh GL, Clemens MA. Lung cancer screening with CT: Mayo Clinic experience. Radiology 2003;226:756-61.

11. Sakurai H, Nakagawa K, Watanabe S, Asamura H. Clinicopathologic features of resected subcentimeter lung cancer. Ann Thorac Surg 2015;99:1731-8.

12. Miller DL, Rowland CM, Deschamps C, Allen MS,

Trastek VF, Pairolero PC. Surgical treatment of non-small cell lung cancer 1 cm or less in diameter. Ann Thorac Surg 2002;73:1545-50; discussion 1550-1.

13. Wu G, Jochems A, Refaee T, Ibrahim A, Yan C, Sanduleanu S, Woodruff HC, Lambin P. Structural and functional radiomics for lung cancer. Eur J Nucl Med Mol Imaging 2021;48:3961-74.

14. Pinsky PF, Gierada DS, Black W, Munden R, Nath H, Aberle D, Kazerooni E. Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. Ann Intern Med 2015;162:485-91.

15. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-10.

16. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.

17. Hatt M, Parmar C, Qi J, El Naqa I. Machine (deep) learning methods for image processing and radiomics. IEEE Transactions on Radiation and Plasma Medical Sciences 2019;3:104-8.

18. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-7.

19. Ren H, Xiao Z, Ling C, Wang J, Wu S, Zeng Y, Li P. Development of a novel nomogram-based model incorporating 3D radiomic signatures and lung CT radiological features for differentiating invasive adenocarcinoma from adenocarcinoma in situ and minimally invasive adenocarcinoma. Quant Imaging Med Surg 2023;13:237-48.

20. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86:2278-324.

21. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Communications of the ACM 2017;60:84-90.

22. Wang D, Zhang T, Li M, Bueno R, Jayender J. 3D deep learning based classification of pulmonary ground glass opacity nodules with automatic segmentation. Comput Med Imaging Graph 2021;88:101814.

23. Yu Y, Wang N, Huang N, Liu X, Zheng Y, Fu Y, Li X, Wu H, Xu J, Cheng J. Determining the invasiveness of ground-glass nodules using a 3D multi-task network. Eur

Radiol 2021;31:7162-71.

24. Ashraf SF, Yin K, Meng CX, Wang Q, Wang Q, Pu J, Dhupar R. Predicting benign, preinvasive, and invasive lung nodules on computed tomography scans using machine learning. J Thorac Cardiovasc Surg 2022;163:1496-1505.e10.

25. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging (Bellingham) 2018;5:011021.

26. Xia X, Gong J, Hao W, Yang T, Lin Y, Wang S, Peng W. Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan. Front Oncol 2020;10:418.

27. Wang X, Li Q, Cai J, Wang W, Xu P, Zhang Y, Fang Q, Fu C, Fan L, Xiao Y, Liu S. Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on CT scan using multi-task learning and deep radiomics. Transl Lung Cancer Res 2020;9:1397-406.

28. Wang C, Shao J, Lv J, Cao Y, Zhu C, Li J, Shen W, Shi L, Liu D, Li W. Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography. Transl Oncol 2021;14:101141.

29. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. Graph neural networks: A review of methods and applications. AI open 2020;1:57-81.

30. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907 2016.

31. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903 2017.

32. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems 2017;30.

33. Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B, Rueckert D. Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease. Med Image Anal 2018;48:117-30.

34. Hao J, Liu J, Pereira E, Liu R, Zhang J, Zhang Y, Yan K, Gong Y, Zheng J, Zhang J, Liu Y, Zhao Y. Uncertainty-guided graph attention network for parapneumonic effusion diagnosis. Med Image Anal 2022;75:102217.

35. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011;12:2825-30.

37. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996;58:267-88.

38. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? arXiv preprint arXiv:181000826 2018.

39. Hu F, Huang H, Jiang Y, Feng M, Wang H, Tang M, Zhou Y, Tan X, Liu Y, Xu C, Ding N, Bai C, Hu J, Yang D, Zhang Y. Discriminating invasive adenocarcinoma among lung pure ground-glass nodules: a multi-parameter prediction model. J Thorac Dis 2021;13:5383-94.

40. Al-Shabi M, Lan BL, Chan WY, Ng KH, Tan M. Lung nodule classification using deep Local-Global networks. Int J Comput Assist Radiol Surg 2019;14:1815-9.

41. Jiang H, Shen F, Gao F, Han W. Learning efficient, explainable and discriminative representations for pulmonary nodules classification. Pattern Recognition 2021;113:107825.