Deep multiple instance learning for predicting chemotherapy response in non-small cell lung cancer using pretreatment CT images

Runsheng Chang¹, Shouliang Qi²,³*, Yanan Wu¹, Qiyuan Song¹, Yong Yue³, Xiaoye Zhang⁴, Yubao Guan⁵ & Wei Qian¹

The individual prognosis of chemotherapy is quite different in non-small cell lung cancer (NSCLC). There is an urgent need to precisely predict and assess the treatment response. To develop a deep multiple-instance learning (DMIL) based model for predicting chemotherapy response in NSCLC in pretreatment CT images. Two datasets of NSCLC patients treated with chemotherapy as the first-line treatment were collected from two hospitals. Dataset 1 (163 response and 138 nonresponse) was used to train, validate, and test the DMIL model and dataset 2 (22 response and 20 nonresponse) was used as the external validation cohort. Five backbone networks in the feature extraction module and three pooling methods were compared. The DMIL with a pre-trained VGG16 backbone and an attention mechanism pooling performed the best, with an accuracy of 0.883 and area under the curve (AUC) of 0.982 on Dataset 1. While using max pooling and convolutional pooling, the AUC was 0.958 and 0.931, respectively. In Dataset 2, the best DMIL model produced an accuracy of 0.833 and AUC of 0.940. Deep learning models based on the MIL can predict chemotherapy response in NSCLC using pretreatment CT images and the pre-trained VGG16 with attention mechanism pooling yielded better predictions.

Globally, lung cancer is the most common cancer and is responsible for the highest incidence and death. Non-small cell lung cancer (NSCLC) is the most common subtype and accounts for 80% of all lung cancers¹. Although the lung cancer death rate has fallen by 54% for men (after 1990) and by 30% for women (after 2002) because of improved healthcare and increased access to early screening, the 2-year and 5-year relative survival rates are still only 36% and 19%, respectively²,³. Furthermore, most of the diagnosed cases are at advanced stages. The individual prognosis for existing clinical treatment options such as surgery, chemotherapy, radiotherapy, targeted drug therapy, and immunotherapy is quite different, and there is an urgent need to precisely predict and assess treatment response⁴.

During clinical treatment, there are significant individual differences in patients with NSCLC. Chemotherapy is the most widely used first-line treatment for lung cancer, and the prediction of its prognostic endpoints will significantly improve clinical applications. No treatment option can completely replace chemotherapy. However, targeted drugs and immune checkpoint inhibitors (as supplements) are gaining increasing importance in treating NSCLC⁵. In clinical practice, oncologists usually judge whether to apply chemotherapy to a patient based on related guides, combined with clinical indicators, biopsy, gene sequencing, and pathological section analysis. The biopsy, gene sequencing, and pathological sectioning procedures are usually invasive. Here, only the local information of the tumor is considered, and the progress of the disease cannot be tracked⁶,⁷.

Radiomics has been widely used to analyze and evaluate medical images⁸–¹⁰. By extracting quantitative features from medical images, radiomics has achieved impressive performance in predicting pathological, radiological, and chemotherapeutic responses in NSCLC¹¹, breast cancer¹², and prostate cancer¹³. However, radiomics has
not yet been translated into clinical practice because it usually requires a selection of the region of interest and manual extraction of features and feature selection\(^{14}\).

In recent years, deep learning models have shown great potential in providing the end-to-end solution from medical images to clinical endpoints\(^{15-18}\). These achievements benefited from the development of deep convolutional neural networks (CNNs). The nonlinear and even high-dimensional information in medical images can be learned automatically by deep CNNs\(^{19-22}\). Xu et al. used recurrent neural network (RNN) to predict survival and cancer-specific outcomes in 179 NSCLC patients with time-series Computed tomography (CT) scans\(^{23}\). Paul et al. used a transfer learning model to extract deep features to predict short- and long-term survivors with lung adenocarcinoma and achieved an accuracy of 90%\(^{30}\).

Multiple instance learning (MIL) is a weakly supervised method and has been widely used in deep learning as it allows for labelling a series of images (bag) instead of each slice (instance)\(^{25}\). This strategy is conducive to extracting global features and can avoid the influence of individual false positive instances\(^{26,27}\). Chen et al. developed a MIL network to extract features in retinopathy images of prematurity to improve staging results\(^{28}\).

Li et al. proposed a multi-resolution MIL model to detect suspicious regions for fine-grained grade prediction in 830 patients\(^{30}\). Li et al. assessed the severity level of COVID-19 on the CT images of 229 cases using a MIL framework\(^{30}\). No MIL studies predicting chemotherapy response in NSCLC have been reported.

In this study, we proposed a deep multiple instance learning (DMIL) model to predict the chemotherapeutic response of NSCLC patients using pretreatment CT images. The contributions contain three primary aspects. Firstly, the proposed DMIL model demonstrates great performance and generalizability for predicting treatment response. Secondly, different pre-trained backbone CNNs are evaluated for feature extraction. Thirdly, an attention mechanism pooling method is proposed and compared with other methods. The developed model is likely to be a weakly supervised, non-invasive, low-cost tool for NSCLC management.

### Materials and methods

#### Patient cohorts.

We enrolled 661 patients from the Shengjing Hospital of China Medical University. The inclusion criteria were that the patient was diagnosed with NSCLC between 2015 and 2019 and treated using chemotherapy alone as the first-line regimen. By excluding patients with no diagnosis reports or no CT images before or after chemotherapy, only 301 patients were included to generate Dataset 1. Using the same inclusion criteria, 42 patients from the Fifth Affiliated Hospital of Guangzhou Medical University were enrolled to generate Dataset 2. This study was approved by the ethics committee of the Shengjing Hospital of China Medical University and the Fifth Affiliated Hospital of Guangzhou Medical University because it was a retrospective study. The characteristics of patients are listed in Table 1, and all experiments were performed in accordance with relevant guidelines and regulations. Waiver for informed consent was approved the ethics committee of the Shengjing Hospital of China Medical University and the Fifth Affiliated Hospital of Guangzhou Medical University. All experiments were performed in accordance with relevant guidelines and regulations. Waiver for informed consent was approved the ethics committee of the Shengjing Hospital of China Medical University and the Fifth Affiliated Hospital of Guangzhou Medical University because it was a retrospective study. The characteristics of patients are listed in Table 1, and the parameters for clinical CT acquisition are listed in Table 2. Dataset 1 was stratified into training (n = 211), validation (n = 30) and test (n = 60) cohorts randomly in a ratio of 70:10:20%. Dataset 2 was used as an external validation cohort to assess the generalizability of the trained models.

The gold standard for measuring the axial diameter of lesions and assessing clinical prognosis is called the response evaluation criteria in solid tumors (RECIST)\(^{31}\). By comparing CT images before and after chemotherapy, all cases were categorized as “response” or “nonresponse.” The response group includes CR (complete response) and PR (partial response), whereas the nonresponse group includes PD (progressive disease) and SD (stable disease). According to RECIST, all tumors were independently assessed by two experienced radiologists, and they reached an agreement through discussion if there was a disagreement. For Dataset 1, the interval between CT scans before and after chemotherapy was 4.74 ± 1.36 (response) and 3.96 ± 2.55 (nonresponse) treatment

| Characteristics | Dataset 1 | Dataset 2 |
|-----------------|-----------|-----------|
| Number of patients | 163 | 22 |
| Gender | Male | 89 | 18 |
| | Female | 74 | 4 |
| Age [years] (Mean ± S. D.) | 63.59 ± 11.28 | 65.86 ± 9.19 |
| Histological type | Adenocarcinoma | 132 | 12 |
| | Squamous cell carcinoma | 31 | 10 |
| Smoking status | Ever | 61 | 16 |
| | Never | 102 | 6 |
| Number of treatment courses | 4.86 ± 1.25 | 1.255 |

Table 1. Clinical characteristics of NSCLC patients. \(^{1}p\) value for the Chi-square test; \(^{2}p\)-value for the two-sample t test.
courses (3-week treatment course), respectively. Whereas, the interval was $3.21 \pm 1.65$ (response) and $3.33 \pm 1.24$ (nonresponse) treatment courses for Dataset 2.

**Model design.** An illustration of the proposed DMIL model is provided in Fig. 1. It consists of three main modules:

1. **Preprocessing module**
   The CT images of NSCLC patients have different slice thicknesses and pixel sizes because of the different acquisition equipment. To eliminate the influence of such variables on the model prediction, we resampled the CT voxels to an isotropic resolution of $1 \times 1 \times 1 \text{ mm}^3$. Considering the accuracy and efficiency of subsequent CNN operations, a cuboid of $64 \times 64 \times 32$ was cropped from CT images to contain the whole tumor. Two experienced radiologists performed this operation after reaching an agreement. A tumor cuboid is treated as a bag, and each patch (or slice) within the cuboid is denoted as an instance. Next, using the grayscale CT images as input to the pre-trained CNN, we converted the images to RGB format and resized each layer to adapt to different pre-trained networks. Each cuboid or bag was marked “response” or “nonresponse.”

2. **Feature extraction module**
   Each instance is input into a path of the model, in which a deep CNN backbone network is used to extract the features. The backbone network is pre-trained by images from ImageNet and fine-tuned on our CT slices. Five baseline pre-trained models (AlexNet, VGG16, ResNet34, DenseNet, and MobileNetV2) are evaluated. The weights in backbone networks are shared across different paths. A mean pooling method was applied to reduce the dimension of extracted features to $32 \times 512$.

3. **Feature representation module**
   Three methods (max pooling, convolutional pooling, and attention mechanism pooling) were used to represent the features of a bag into a fully connected layer and then classify it into response and nonresponse categories. Through the combination of five pre-trained backbone networks with transfer learning and three pooling methods, several models were obtained and compared.

**Neural network structure.** Pre-trained models like VGG16 are widely used to extract features from images because of their simple structure and good generalization performance in transfer learning. All weights were trained in the set of 14 million 2D color images called ImageNet. The network structure was implemented in Python 3.9, using PyTorch 1.10.

| Parameter                      | Dataset 1       | Dataset 2       |
|-------------------------------|-----------------|-----------------|
| kVp [kV]                      | 120             | 120             |
| X-ray tube current (mean ± S.D.) [mA] | 226.24 ± 66.86  | 254.25 ± 54.48  |
| Slice thickness [mm]           | 2.5 (n = 19); 3.0 (n = 262); 5.0 (n = 20) | 2.0 (n = 42) |
| Pixel size [mm]                | 0.74 ± 0.15     | 0.85 ± 0.67     |
| CT scanner manufacturer        | GE Medical (n = 12), Siemens (n = 12), Toshiba (n = 15), Philips (n = 262) | Siemens (n = 42) |

**Table 2.** Parameters for CT image acquisition.
Figure 2 shows that VGG16 has 5 convolutional layers, each segment has 2 or 3 convolutional layers, and a maximum pooling layer is connected at the end of each segment to reduce the image size. The number of convolutional kernels in each segment is the same, and the closer to the fully connected layer, the more convolutional kernels there are. The last segment is a fully connected network, and we added a pooling segment to represent features before the fully connected segment.

The specific process of network structure is given as follows. Firstly, each CT slice from one patient (32 × 224 × 224) was passed (in order) through the feature extractor based on VGG16. A 32 × 512 × 7 × 7 tensor was then obtained, and it contained features for each slice. Secondly, a global average pooling layer was applied to reduce the tensor dimension to 32 × 512. Thirdly, we input this tensor into the feature representation module. Three methods were used to represent the features of a bag:

1. **Max pooling**
   A max pooling was applied across slices to obtain a 512-dimensional vector, and this operation was performed between slices. Among the 512 kernels containing features, the largest feature in every 32 slices is taken to represent one feature kernel. Then it is passed to a fully connected layer, and the sigmoid activation function is used to obtain a prediction in the 0 to 1 range (Fig. 3a).

2. **Conv pooling**
   A 32 × 1 convolutional layer was applied to this tensor and obtained a 512-dimensional vector. This convolutional layer is applied to obtain a feature vector that can characterize each convolution kernel using a convolution operation (Fig. 3b).

3. **Attention mechanism pooling**
   This DMIL design is commonly referred to as “Attention-based multiple instance learning”, as presented in the study by Ilse et al. Two fully connected layers were applied to obtain a 32 × 1 vector called attention weight and then it was multiplied by the 32 × 512 tensor to obtain a 512-dimensional vector (Fig. 3c). This attention mechanism pooling determines the weights to instances by a neural network. Additionally, the sum of all weights must be equal to 1 for a bag-level prediction. The weighted average meets the requirements that the weights and the embedding are part of the \( f \) function. Let \( H = \{h_1, \ldots, h_k\} \) be a bag of \( k \) embeddings, and then we propose the following MIL pooling:

\[
    z = \sum_{k=1}^{k} a_k h_k
\]  

where
Statistical analysis. Performance measures for each model included the accuracy, sensitivity, specificity, F1-score, the area under the receiver operating characteristic curve (AUC), and confusion matrix. The decision curve analysis (DCA) was also used to evaluate each model by quantifying the benefit at different threshold probabilities. We also provided a 95% confidence interval (CI) for AUC, the cut-off using Youden’s index, and the shortest distance from the coordinate (0, 1) for the ROC curve to assess the variability in estimates.

To assess whether these findings were dependent on the clinical features, we performed a two-sample t-test to compare the age and number of treatment courses and a chi-square test to compare the gender, histological type, and smoking status between the response and nonresponse groups. The significance level was set to 5%.

Results

Clinical characteristics. The information on clinical characteristics is shown in Table 1. In Dataset 1, there were no significant differences between the response and nonresponse groups for the measures of gender, age, histological type, smoking status, and the number of treatment courses. Similarly, no significant differences were observed between the two groups in Dataset 2.

Performance of the DMIL model. The performance of the DMIL model is shown in Fig. 4 and Table 3. As the radar chart demonstrates in Fig. 4a, the accuracy, AUC, sensitivity, specificity, and F1-score were 0.883, 0.982, 0.871, 0.867, and 0.885, respectively, whereas the cut-off value was 0.833. Figure 4b,c show the ROC and confusion matrix. The model correctly predicted 27 of 30 nonresponse and 26 of 30 response patients in the test cohort. According to the DCA shown in Fig. 4d, the DMIL curve showed much more net benefit than the baseline curves of “Treat none” and “Treat all”.

Comparison with the counterparts. Five different pre-trained models were investigated in the feature extraction module while using the attention mechanism pooling method. The performance of the four counterparts is shown in Fig. 5 and Table 3. The accuracy, AUC, sensitivity, specificity, and F1-score are shown in the radar chart (Fig. 5a). The ROC and confusion matrix are shown in Fig. 5b,c, the AUC of AlexNet, ResNet34,
DenseNet and MobileNet_v2 were 0.951, 0.864, 0.927 and 0.963, respectively. The DMIL model had a higher AUC than the four backbone models. This finding was confirmed by DCA (Fig. 5d) and proves that the DMIL model is more robust than the four backbone models.

Three pooling methods combined with the VGG16 model were studied in the feature representation module. The performance of the two counterparts is shown in Fig. 6 and Table 3. The accuracy, AUC, sensitivity, specificity, and F1-score of the DMIL models with different backbone networks are presented in Table 3.

Table 3. Performance of the DMIL models with different backbone networks. *AM attention mechanism, ACC accuracy, AUC area under the curve, SEN sensitivity, SPE specificity.

DenseNet and MobileNet_v2 were 0.951, 0.864, 0.927 and 0.963, respectively. The DMIL model had a higher AUC than the four backbone models. This finding was confirmed by DCA (Fig. 5d) and proves that the DMIL model is more robust than the four backbone models.

Three pooling methods combined with the VGG16 model were studied in the feature representation module. The performance of the two counterparts is shown in Fig. 6 and Table 3. The accuracy, AUC, sensitivity, specificity, and F1-score of the DMIL models with different backbone networks are presented in Table 3.

Figure 4. Performance of the DMIL model in the test cohort. (a) Statistical results; (b) The ROC curve; (c) The confusion matrix; (d) The DCA result.
specificity, and F1-score are shown in the radar chart of Fig. 6a. The ROC and confusion matrix are shown in Fig. 6b,c, and the AUC of max pooling and convolutional pooling methods were 0.958 and 0.931, respectively. Compared with the DMIL model, both pooling methods had a lower AUC. In the DCA (Fig. 6d), the DMIL model had better performance than both pooling methods.

During the model training, the Adam optimizer was utilized. The learning rate, batch size, and the number of epochs were set as 10–5, 32, and 50, respectively.

Performance in the external validation dataset. In the external validation cohort (Dataset 2), 42 NSCLC patients were used to validate the DMIL model, and the performance is shown in Fig. 7. As the radar chart shows in Fig. 7a, the accuracy, AUC, sensitivity, specificity, and F1-score were 0.833, 0.940, 0.842, 0.864, and 0.821, respectively, using a cut-off value of 0.805. The ROC and CM are shown in Fig. 7b,c, and the model correctly predicted 17 of 20 nonresponse and 18 of 22 response patients. The DCA results in Fig. 7d indicated that the curve of the DMIL model was higher compared with the treat none and treat all groups.

Discussion
In this study, a deep learning model based on MIL was proposed to predict chemotherapy response in NSCLC patients using CT images. The model with pretrained VGG16 backbone and an attention mechanism pooling method demonstrated the best performance, comparing 15 different designs. Furthermore, good generalizability was observed when using an external validation dataset.

A weakly supervised method called multiple instance learning was used to train the model. Each patient in the dataset is regarded as a bag, and each bag contains all slices of one patient, which is regarded as an instance. Each bag has a training label, and the instances in the bag are unlabeled. A bag is given a positive label if at least one positively-labeled instance is in the bag. For a negatively-labeled bag, all its instances are negatively-labeled. The advantage of this weakly supervised learning method over supervised learning is that not every CT slice of the patient is adequate for evaluating the efficacy of chemotherapy. Training labels based on the bag instead of instances can eliminate the noise that has no predictive effect on the chemotherapy effect and reduces the false positives in the training process.

Using pre-trained models as the backbone module for computer vision and natural language processing tasks is popular in deep learning. Because the development of neural network models for these tasks requires a lot of
computational resources and time, pre-trained models can considerably boost related works. There are three main advantages to pre-trained models: (a) better initialization: the initial skill in the source model is higher; (b) faster convergence: the rate of improvement in skill is steeper during training of the source model than in other cases; (c) higher asymptote: the fusion skill of the trained models is better than the other cases. We used five different pre-trained models (AlexNet, VGG16, ResNet34, DenseNet, and MobileNet_v2) as the backbone module to extract features from CT slices. We found that VGG16 had better predictive performance than the other models. This model is characterized by stacking several convolutional layers and pooling layers, and the convolutional layers and pooling layers both use the same convolution kernel and pooling kernel parameters, which can easily form a deeper network structure. Overall, VGG16 has small filters and deeper networks.

We tried three strategies to represent features in the pooling module: max pooling, convolutional pooling, and attention mechanism pooling. For the max pooling method, the features of one bag are represented by taking the maximum value of all layers in each convolutional kernel. This may lose important information that is not the maximum value. For the convolutional pooling method, the information of all slices in one bag is calculated to represent global features. However, this single-layer convolution is not fed back to the network before, so the critical information may also be lost. The attention mechanism pooling method combines a fully connected and attention mechanism. Using two fully connected layers, this method can obtain the weights of all layers for each convolution kernel. After this, multiplying the weights with the previous features produces a matrix representing the global features. This feedback mechanism can update and optimize the network parameters while preserving the features of each slice.

After further validation, this DMIL model might be applied to clinical practice. Firstly, clinical oncologists can make personalized treatment plans for patients according to the predicted possibilities of response or nonresponse to the chemotherapy. If the possibility of nonresponse is very high, combined chemotherapy, radiotherapy, immunotherapy, or other programs should be considered with priority. Secondly, this model is non-invasive and low-cost by utilizing the prechemotherapy CT images scanned in clinical routines. It can be extended to patients who are of low economic affordability and cannot tolerate invasive measurements. Thirdly, this model is proven to have good generalizability, indicating that different hospitals and regions can introduce the model without fine-tuning.

Despite the excellent performance, there are some limitations in this research. Firstly, the five models applied in this study were all commonly used baseline architectures. More modern architectures should be tested with our proposed DMIL design in future studies. Secondly, the slices of CT images input to the model were clipped. We could try raw CT images as the input to see if it benefits the performance in future studies. Thirdly, unlike

Figure 6. Performance of the two counterparts with different representation modules. (a) Statistical results; (b) The ROC curve; (c) The confusion matrix; (d) The DCA result.
radiomic features, the features obtained by the feature representation module were unexplainable because the features extracted by VGG-16 from all instances have been fused and represented once again.

Conclusion
Deep learning models based on MIL can predict chemotherapy response in NSCLC using pretreatment CT images. In this model, the pre-trained VGG16 through transfer learning can efficiently extract features from each instance, and an attention mechanism pooling module can estimate the importance of different instances and represent features of the bag. These excellent feature extraction and representation methods enable good model performance and generalization ability. After further clinical validation, the developed model is likely to be a weakly supervised, non-invasive, low-cost tool for NSCLC management.

Data availability
The datasets used in this study are available on reasonable request to Shouliang Qi (qisl@bmie.neu.edu.cn).

Received: 29 May 2022; Accepted: 14 November 2022
Published online: 18 November 2022

References
1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2021. CA Cancer J. Clin. 71(1), 7–33 (2021).
2. Liu-Jarin, X. Histologic assessment of non-small cell lung carcinoma after neoadjuvant therapy. Mod. Pathol. 16(11), 1102 (2003).
3. Fried, D. V. et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int. J. Radiat. Oncol. Biol. Phys. 90(4), 834–842 (2014).
4. Wang, M., Herbst, R. S. & Boshoff, C. Toward personalized treatment approaches for non-small-cell lung cancer. Nat. Med. 27(8), 1345–1356 (2021).
5. Pirker, R. Chemotherapy remains a cornerstone in the treatment of nonsmall cell lung cancer. Curr. Opin. Oncol. 32(1), 63–67 (2020).
6. Gadgeel, S. M. Role of chemotherapy and targeted therapy in early-stage non-small cell lung cancer. Am. Soc. Clin. Oncol. Educ. Book 37(1), 630–639 (2017).
7. Dafni, U., Tsourti, Z., Vervita, K. & Peters, S. Immune checkpoint inhibitors, alone or in combination with chemotherapy, as first-line treatment for advanced non-small cell lung cancer. A systematic review and network meta-analysis. Lung Cancer 134, 127–140 (2019).
8. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. Nat. Rev. Cancer. 18(8), 500–510 (2018).

Figure 7. Performance of the DMIL model in the external validation. (a) Statistical results; (b) The ROC curve; (c) The confusion matrix; (d) The DCA result.
11. Lambin, P .
14. Chetan, M. R. & Gleeson, F . V . Radiomics in predicting treatment response in non-small-cell lung cancer: Current status, challenges and future perspectives. Eur. Radiol. 31(2), 1049–1058 (2020).
15. Parmar, C., Barry, J . D ., Hosny, A ., Quackenbush, J . & Aerts, H . Data analysis strategies in medical imaging. Clin. Cancer Res. 24(15), 3492–3499 (2018).
16. Coroller, T . P . et al . Radiomic-based pathological response prediction from primary tumors and lymph nodes in NSCLC. J. Thorac. Oncol. https://doi.org/10.1016/j.jtho.2016.11.2226 (2016).
17. Coroller, T . P . et al . Radiomic phenotype features predict pathological response in non-small cell lung cancer. Radiother. Oncol. 119(3), 480–486 (2016).
18. Coroller, T . P . C . A . et al . CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother. Oncol. 114(3), 345–350 (2015).
19. Shin, H . C . et al . Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35(5), 1285–1298 (2016).
20. Setio, A . A . A . et al . Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging 35(5), 1160–1169 (2016).
21. Dey, N ., Zhang, Y . D ., Rajinikanth, V ., Pugalenthi, R . & Raja, N . Customized vgg19 architecture for pneumonia detection in chest X-rays. Pattern Recognit. Lett. 143, 67–74 (2021).
22. Souid, A ., Sakli, N . & Sakli, H . Classification and predictions of lung diseases from chest X-rays using MobileNet V2. Appl. Sci. 11(6), 2751 (2021).
23. Xu, Y . et al . Deep learning predicts lung cancer treatment response from serial medical imaging. Clin. Cancer Res. 25(11), 3266–3275 (2019).
24. Paul, R . et al . Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomogr. A J. Imaging Res. 2(4), 388–395 (2016).
25. Dietterich, T ., Lathrop, R . H . & Lozano-Pérez, T . Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89(1–2), 31–71 (1997).
26. Maron, O . & Lozano-Pérez, T . A framework for multiple-instance learning. In Neural Information Processing System (NIPS) 10 (MIT, 1997).
27. Chen, Y ., Bi, J . & Wang, J . Z . MILES: Multiple-instance learning via embedded instance selection. IEEE Trans. Pattern Anal. Mach. Intell. 28(12), 1931 (2006).
28. Chen, S ., Zhang, R ., Chen, G ., Zhao, J . & Lei, B . Attention-guided deep multi-instance learning for staging retinopathy of prematurity. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021).
29. Li, J ., Li, W ., Sisk, A ., Ye, H . & Arnold, C . W . A multi-resolution model for histopathology image classification and localization with multiple instance learning. Comput. Biol. Med. 131(9), 104253 (2021).
30. Li, Z . et al . A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. Med. Image Anal. 69, 101978 (2021).
31. Eisenhauer, E . A . et al . New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). EUR J CANCER 45(2), 228–247 (2009).
32. Deng, J . et al . ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (IEEE, 2009).
33. Krizhevsky, A ., Sutskever, I . & Hinton, G . ImageNet classification with deep convolutional neural networks. Communications of the ACM. 60(6), 84–90 (2017).
34. Huang, G ., Liu, Z ., Laurens, V . & Weinberger, K . Q . Densely connected convolutional networks. In IEEE Computer Society (2016).
35. Simonyan, K . & Zisserman, A . Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
36. Sandler, M ., Howard, A ., Zhu, M ., Zhmoginov, A . & Chen, L . C . MobileNetV2: Inverted residuals and linear bottlenecks. In IEEE Conference on Computer Vision and Pattern Recognition, 2019, 109(1), 43–76 (2020).
37. He, K ., Zhang, X ., Ren, S . & Sun, J . Deep residual learning for image recognition In IEEE (2016).
38. Bien, N ., Rajpurkar, P ., Ball, R . L ., Irvin, J . & Lungren, M . P . Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med. 15(11), e1002699 (2018).
39. Moranguiño, J ., Pereira, T ., Ramos, B ., Morgado, J ., Costa, J . L . & Oliveira, H . P . Attention based deep multiple instance learning approach for lung cancer prediction using histopathological images. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2852–2855 (IEEE, 2021).
40. Paszke, A . et al . Pytorch: An imperative style, high-performance deep learning library. In Neural Information Processing System (NIPS) 32 (MIT, 2019).
41. Ilse, M ., Tomczak, J . & Welling, M . Attention-based multiple instance learning. In International Conference on Machine Learning, 2127–2136 (PMLR, 2018).
42. Han, Z . et al . Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. IEEE Trans. Med. Imaging 39(8), 2584–2594 (2020).
43. Cholgyoina, V ., Tax, D . M . & Loog, M . Multiple instance learning with bag dissimilarities. Pattern Recognit. 48(1), 264–275 (2015).
44. Zhuang, F . et al . A comprehensive survey on transfer learning. Proc. IEEE 109(1), 43–76 (2020).
45. Tan, C ., Sun, F ., Kong, T ., Zhang, W ., Yang, C . & Liu, C . A survey on deep transfer learning. In International Conference on Artificial Neural Networks, 270–279 (Springer, 2018).
46. Raffel, C . et al . Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019).
47. Raghu, M ., Zhang, C ., Kleinberg, J . & Bengio, S . Transferfusion: Understanding transfer learning for medical imaging. In Neural Information Processing System (NIPS) 32 (MIT, 2019).
48. Niu, Z ., Zhong, G . & Yu, H . A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62 (2021).
49. Bello, I ., Zoph, B ., Vaswani, A ., Shlens, J . & Le, Q . V . Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 3286–3295 (2019).
50. Ramachandran, P . et al . Stand-alone self-attention in vision models. In Neural Information Processing System (NIPS) 32 (MIT, 2019).
51. Vaswani, A . et al . Attention is all you need. In Neural Information Processing System (NIPS) 30 (MIT, 2017).
52. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. & Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019).
53. Chen, A., Saouaf, J., Zhou, B., Crawford, R. & Goldmacher, G.A Deep learning-facilitated radiomics solution for the prediction of lung lesion shrinkage in non-small cell lung cancer trials. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (2020).
54. Xu, Y. et al. "Deep learning predicts lung cancer treatment response from serial medical imaging," (in eng). Clin. Cancer Res. 25(11), 3266–3275. https://doi.org/10.1158/1078-0432.Ccr-18-2495 (2019).

Acknowledgements
This work was partly supported by the National Natural Science Foundation of China under Grant (Nos. 82072008), Natural Science Foundation of Liaoning Province (2021-YGJC-C1, 2020-BS-049), Key R&D Program Guidance Projects in Liaoning Province (2019JH8/10300051), and the Fundamental Research Funds for the Central Universities (N2119070, N2224001-10).

Author contributions
R.C. performed experiments and analyzed the data with Y.W. and Q.S., S.Q., X.Z., and W.Q. proposed the idea, made discussions, and composed the manuscript together with R.C. Y.Y., X.Z., and Y.G. collected and analyzed the data. All authors have read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022