Clinical and Translational Science

ARTICLE OPEN ACCESS

MoLPre: A Machine Learning Model to Predict Metastasis of cT1 Solid Lung Cancer

Jie Lan[1] | Heng Wang[1] | Jing Huang[2] | Weiyi Li[2] | Min Ao[2] | Wanfeng Zhang[1] | Junhao Mu[2] | Li Yang[2] | Longke Ran[1,3]

[1]Department of Bioinformatics, The Basic Medical School of Chongqing Medical University, Chongqing, China | [2]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China | [3]Department of Oncology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Correspondence: Li Yang (204534@hospital.cqmu.edu.cn) | Longke Ran (ranlongke@cqmu.edu.cn)

Received: 4 September 2024 | **Revised:** 29 December 2024 | **Accepted:** 10 January 2025

Funding: This work was supported by the Program for Youth Innovation in Future Medicine, Chongqing Medical University (Grant No. W0102) and The National Natural Science Foundation of China (Grant No. 82203181).

Keywords: cT1 solid lung cancer | machine learning | prediction model | pulmonary nodules | random forest

ABSTRACT

Given that more than 20% of patients with cT1 solid NSCLC showed nodal or extrathoracic metastasis, early detection of metastasis is crucial and urgent for improving therapeutic planning and patients' risk stratification in clinical practice. This study collected clinicopathological variables from the pulmonary nodule and lung cancer database of the First Affiliated Hospital of Chongqing Medical University, where patients with early-stage (cT1) solitary lung cancer were evaluated from 2018.11 to 2022.10. The random forest model and Shapley Additive Explanations (SHAP) were used to investigate the importance of clinical features in the feature selection part. Random Forest, Gradient Boosting, and AdaBoost classifiers were applied to build the final model, and the predictive discrimination of each model was compared based on the receiver operating characteristics (ROC) curve and precision and recall curve. With the evaluation of feature importance, 9 features were used to construct the prediction model finally. The Random Forest model yielded an average precision of 0.93 with an area under the curve (AUC) of 0.92 (95% CI: 0.88–0.94) compared with the Gradient Boosting and AdaBoost classifiers in the internal validation dataset, yielding an average precision of 0.87 and 0.91 with AUCs of 0.87 (95% CI: 0.84–0.93) and 0.90 (95% CI: 0.86–0.92), respectively. In addition, the Random Forest classifier performed best in 5 other 5 diagnostic indices. Furthermore, we embedded this model in a web application called MoLPre (https://molpre.cqmu.edu.cn/), a user-friendly tool assisting in the metastasis prediction of cT1 solid lung cancer.

JEL Classification: Artificial Intelligence and Machine Learning

1 | Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, with 1.8 million cases in 2022, despite significant treatment advancements [1]. Owing to the use of low-dose computed tomography (LDCT), an increasing number of small solid lung cancers have been detected [2]. Tumors smaller than 30 mm were classified as clinical (cT1) stage according to the 8th edition of the American Joint Committee on Cancer (AJCC). It is well known that cancer cell metastasis is the primary reason responsible for unsatisfactory prognosis [3]. Tumor size is considered to have a significantly associated with metastasis [4, 5]. However, some patients with malignant solid lung nodules (diameter ≤ 30 mm) showed distal metastases during

Jie Lan and Heng Wang contributed equally to this study.

Summary

- What is the current knowledge on the topic?
  - Even though the number of small-sized solid lung cancers is increasing, tumor size is still regarded as an important risk factor for metastasis. Thus, early metastasis detection of cT1 solid lung cancer is crucial and urgent.
- What question did this study address?
  - More than 20% of patients with cT1 solid NSCLC showed nodal or extrathoracic metastasis. However, current efforts aim to develop machine learning models for the metastasis prediction of lung cancer, ignoring the necessity of tool development for cT1 solid lung cancer.
- What does this study add to our knowledge?
  - This study assesses 16 covariates as predictors of metastasis using ML. This is the first prediction model based on the patients with cT1 solid lung cancer.
- How might this change clinical pharmacology or translational science
  - cT1 solid lung cancer is the major origin of advanced lung cancer and needs to be considered with more closely surveillance and systemic treatment. This study suggested that MoLPre is good for adequate treatment and can reduce morbidity of patients with cT1 solid lung cancer.

diagnosis. Studies have shown that the percentage of nodal or extrathoracic metastases is >20% in cT1 non-small cell lung cancer (NSCLC) patients [6, 7]. Small cell lung cancer (SCLC) manifests with more aggressive characteristics and higher metastatic potential. Only 5% of SCLC cases are early stage when first diagnosed [8]. This kind of solid cT1 lung cancer is the major origin of advanced lung cancer and needs to be considered with closer surveillance and systemic treatment. Therefore, the early detection of metastasis is crucial for improving therapeutic planning and patient risk stratification in clinical practice. Positron emission tomography (PET) with the glucose analog 2-fluoro-2-deoxyglucose (18F-FDG) takes advantage of the high glucose metabolism of lung cancer cells and metastatic lesions to visualize not only tumors but also local lymph node metastasis and other distant metastases. Nevertheless, the misdiagnosis and the false-negative rate remain a matter of concern [9, 10]. The accuracy of PET/CT in detecting lymph node metastasis is low, especially in tuberculosis endemic countries. Endobronchial ultrasound (EBUS)-guided biopsy and thoracoscopy for evaluating mediastinal lymph nodes are invasive procedures [11]. In addition, the serum carcinoembryonic antigen (CEA), recombinant cytokeratin fragment antigen (CYFRA), neuron-specific enolase (NSE), and carbohydrate antigen (CA) series are traditional and common tests used to assist in the diagnosis of tumors, and some studies have shown that the combination of these biomarkers could improve diagnostic ability [12, 13]. Environmental exposures, such as indoor dust, are not only associated with pulmonary inflammation but also promote lung cancer metastasis by inducing tumor necrosis factor-α [14]. It is important to construct a careful evaluation tool to identify metastases before invasive procedures to minimize surgical risk and reduce unnecessary function loss.

With the rapid development of computer science, several deep learning methods have been introduced for medical analysis to perform various tasks, such as cancer diagnosis, malignant probability prediction, and metastasis status classification [15, 16]. The development of quantitative imaging methods, along with machine learning, has helped researchers interpret and extract imaging information [17]. The use of CT/PET imaging can augment patient stratification, prognosis, and prediction. Therefore, a number of models have been developed based on machine learning using high-throughput imaging. Tobias et al. developed a Gradient Boosting classifier to improve the prediction of mediastinal lymph node metastases in NSCLC using 18F-FDG PET/CT parameters [18]. Wang et al. proposed a machine learning-based model named PKU-M, which was developed based on the boosted ensemble algorithm (XGBoost) to estimate the probability of malignancy for multiple pulmonary nodules (MPNs). External validation showed that the PKU-M model was excellent in the discrimination of MPNs, with an AUC of 0.89 (95% CI: 0.86–0.92) [19]. Zhang et al. built a logistic model with nomograms as a convenient and valuable tool for risk evaluation of metastasis in different regions of clinical stage I NSCLC. Validation of the hilar-intrapulmonary node metastasis (HNM) and the mediastinal node metastasis (MNM) in patients got AUCs of 0.87 (95% CI: 0.83–0.91) and 0.82 (95% CI: 0.77–0.88), respectively [20]. Zeng et al. reported a regression model based on the serum biomarkers described previously, and the model was validated to be valuable for assessing tumor metastasis in lung cancer patients [21].

A number of researchers have aimed to assess the possibility of metastasis in lung cancer, but only limited information is available on metastasis for cT1 solid lung cancer. There is a clinical need for new, robust, cost-effective, and convenient non-invasive methods to better predict the metastatic status of cT1 solid lung cancer because incorrect assessments may lead to delayed diagnosis and increase the risk of complications. In this study, we sought to analyze real-world data on the clinical characteristics of the cT1solid lung cancer patients with or without metastases and developed a machine learning model to predict metastasis, providing guidance in decision-making for individualized therapy.

2 | Methods

This study was conducted in the Department of Bioinformatics at the Basic Medical School and the Department of Respiratory and Critical Care Medicine at the First Affiliated Hospital of Chongqing Medical University in China. All patients provided written informed consent prior to enrollment. This study was reviewed and approved by the Ethics Review Committee of the First Affiliated Hospital of Chongqing Medical University.

2.1 | Patients Enrollment

Our study collected all relevant data from the pulmonary nodule and lung cancer database of the First Affiliated Hospital of Chongqing Medical University, and all patient information was enrolled ranging from 2018.11 to 2022.10. This dataset encompasses information on a total of 418 patients who have undergone

rigorous screening and have been diagnosed with early-stage (cT1) solid lung cancer, consisting of 213 non-metastatic and 205 metastatic patients.

The inclusion criteria were as follows: (a) the size of the pulmonary nodules was no more than 30 mm; (b) the pulmonary nodule was solitary; (c) patients were diagnosed with lung cancer, which was assigned according to the AJCC 8th edition; (d) patients with complete basic clinical information, such as age, sex, smoking, exposure, and tumor history; (e) patients with any null values of the five biomarkers were excluded (CEA, NSE, SCC, Pro-GRP, and CYFRA21-1); (f) patients took part in pulmonary nodules and lung cancer whole-course management at the First Affiliated Hospital of Chongqing Medical University.

For patient status identification, the process of diagnosis for metastatic lung cancer was as follows: (a) imaging methods (CT, PET-CT, or ultrasound); (b) a three-level hierarchical medical system; (c) evaluation after anti-tumor treatment; or (d) pathological specimens.

After applying the inclusion and exclusion criteria, 148 cases with non-metastatic nodules and 138 cases with metastatic nodules remained for the construction of the prediction model.

2.2 | Study Design

Figure 1 presents an overview of the workflow used in the study. All information about patients with pulmonary nodules was collected, and patients were randomly separated into training and validation datasets at a ratio of 7:3. A grid search through 10-fold cross-validation was performed to ascertain the best parameter combination value of the model; the classifier models were trained and then evaluated for performance in the internal validation dataset. In addition, the final model was compared with the two models in terms of precision, AUC, accuracy, etc.

2.3 | Outcome and Prediction Variables

All patients in this study were labeled as non-metastatic or metastatic, with or without clinically node-negative tumors or distal metastasis. The outcome of interest was metastasis to pulmonary nodules, including lymph nodes or distal metastases. The variables of interest were clinical and pathological features, age, sex, smoking history, tumor history, environmental exposures, nodule number, nodule location, the largest nodule dimension, and five tumor biomarkers (CEA, NSE, SCC, Pro-GRP, and CYFRA21-1).

2.4 | Model Training and Validation

To ensure the objectivity and reliability of model assessments, we randomly split the data at a ratio of 7:3 as training and validation datasets. Next, as there were 15 features in all, some features may not be important for the clinical outcome. To select important features in an unbiased manner, we investigated the importance of all the features using the Random Forest model. Indeed, SHAP value analysis was applied to

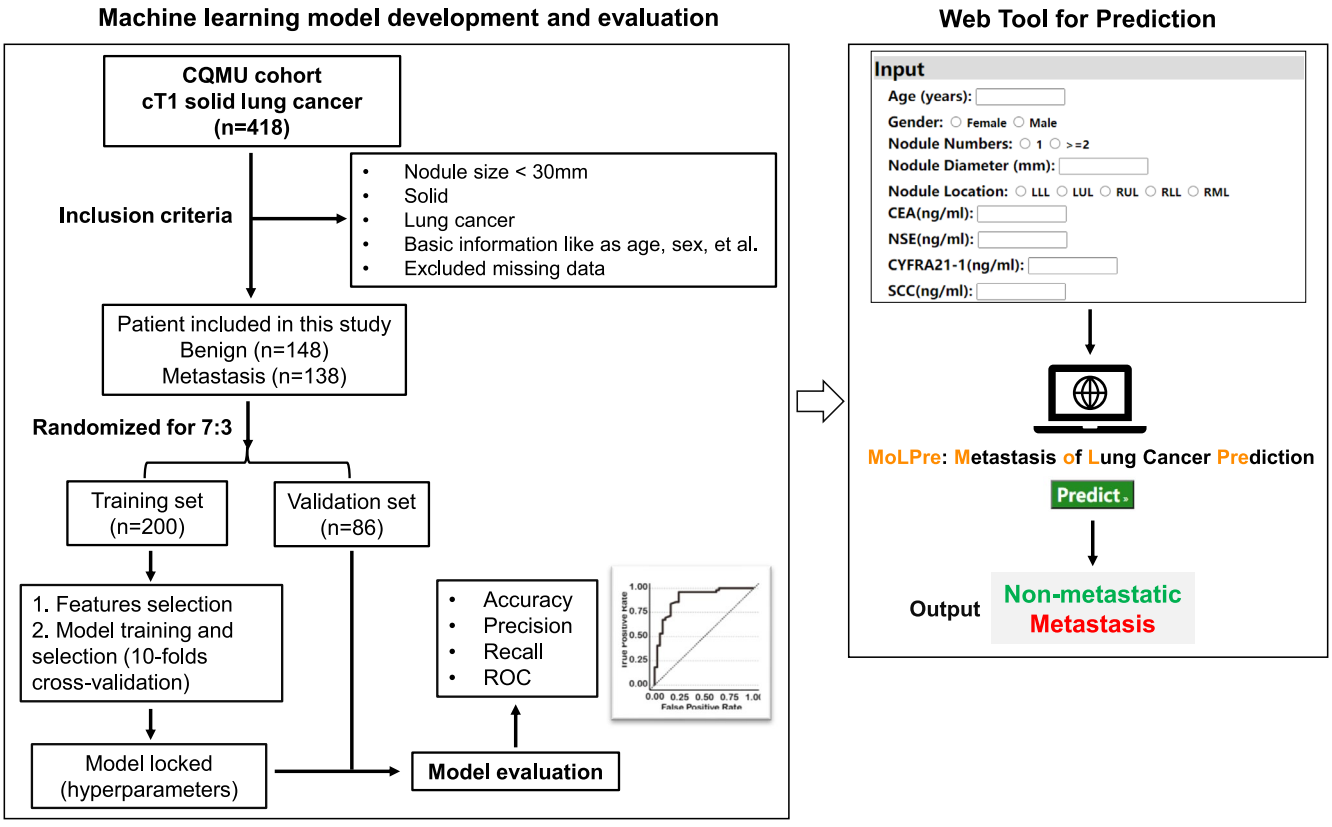

FIGURE 1 | Study pipeline. CQMU: Chongqing medical university. ROC: Receiver Operating Characteristics. Abbreviations: CEA, Carcinoembryonic antigen; CYFRA21-1, Recombinant cytokeratin fragment antigen 21–1; NSE, Neuron-specific enolase; CA, Carbohydrate antigen.

understand how each feature contributes to the model's output and how they affect the final prediction [22, 23]. By repeating the 10-fold cross-validation through a grid search along numerous sets of hyperparameters, we obtained the best hyperparameters. In our Random Forest model, the hyperparameters were set as follows: criterion, entropy; maximum depth of tree, 9; maximum features, log2; minimal samples in a leaf, 8; minimal samples in tree split, 3; number of tree estimators, 100. For the Gradient Boosting model, we set the learning rate to 0.01, loss of function to exponential, number of tree estimators to 80, and subsample rate to 0.9. For the AdaBoost model, we set the algorithm to SAMME, learning rate to 0.7, and number of tree estimators to 100. Furthermore, with pairwise Pearson correlation analysis with Benjamini-Hochberg (BH) adjustment of these five features, we chose the top nine features (age, sex, nodule number, nodule location, largest nodule dimension, CEA, NSE, SCC, and CYFRA21-1) to incorporate into our final prediction model. Next, using these nine features, we trained three classifiers in the training dataset, and then applied the classifiers to the validation dataset Classifier performances were assessed based on their precision, AUC, and other 5 diagnostic indices. In terms of the 95% CIs, we employed the bootstrap method to conduct 10 sampling iterations, and 70% of the validation dataset was used for model evaluation.

2.5 | Survival Analysis

To assess the impact of clinical characteristics on the prognosis of patients with cT1 solid lung cancer, we will perform survival analyses. Nine characteristics will be included: CEA, nodule diameter, CYFRA21-1, NSE, SCC, age, nodule number, nodule location, and sex.

2.6 | Implementation

For user-friendly access, the final model was also implemented as a web-based tool, and users could predict metastasis by filling in the patient and nodule characteristics without registration. NumPy (version:1.23.5), Pandas (version:1.5.1), and Scikit-learn (version:1.1.3) were used to build the prediction model. The pickle module was used to package our model, which is available in our web portal (https://molpre.cqmu.edu.cn/). The SHAP module was used to interpret the classifier. The web portal was developed using HTML and CSS scripts and implemented in Python based on the Django web framework. All the backend scripts were written in the Python programming language as well.

2.7 | Statistical Analysis

All features were analyzed using the Wilcoxon rank sum test or Fisher's exact test to test the correlations between non-metastatic and metastatic patients. Recurrence and survival probabilities were estimated using the Kaplan–Meier method and compared using the log-rank test. The ROC curve and corresponding AUC were applied to investigate the performance of different classifier models, and the 95% CI was calculated by 10 random samplings with 70% of the validation samples. Statistical analyses were performed using R statistics software (version:3.6.0; http://www.r-project.org/) with R packages ggplot2 (version:3.3.3), psych (version:2.1.6), corrplot (version:0.89), survival (version:3.2.11), survminer (version:0.4.9), and reshape2 (version:1.4.4).

3 | Results

3.1 | Patients and Clinical Characteristics

A total of 286 patients were included in the study, and nearly all the patients were found to have either local or distant metastasis when diagnosed with malignant lung tumors. The corresponding clinical characteristics of the two datasets are described in Table 1. The study included 127 women (44.4%) and 159 men (55.6%) with a mean (SD) age of 62.91 (10.90) years. A total of 138 patients (48%) had pulmonary nodules with lymph node or distal metastasis, and there was no evidence of metastasis in the remaining patients (52%). Patients with metastasis showed a significantly larger nodule diameter and higher levels of CEA, NSE, and CYFRA21-1 (all $p < 0.05$). In addition, nodule numbers and histological classification showed significant differences between patients with and without metastasis (all $p < 0.05$). The percentage of patients with no fewer than two nodules in the metastasis group was 67%, and the percentage of patients with no fewer than two nodules in the non-metastatic group was 42%. However, age, sex, smoking status, environmental exposure, history of family tumor, prior tumor, location of nodule, micropapillary, Pro-GRP, and SCC were not significantly different between patients with and without metastasis. All patients were randomly separated into training and validation datasets containing 200 and 86 patients, respectively (Figure 1). All features showed no significant differences between the training and validation datasets by Fisher's exact test or Wilcoxon rank sum test (Figure S1).

3.2 | Feature Selection and Cross-Validation

We investigated feature importance using the Random Forest and SHAP analysis, as summarized in Figure 2A,C. The importance evaluation results indicated that CEA was the most important feature, followed by nodule diameter, Pro-GRP, CYFRA21-1, and NSE. With pairwise Pearson correlation analysis with BH adjustment for these five biomarkers, Pro-GRP showed a significant positive correlation with NSE (Figure 2B). However, Pro-GRP did not show a significant difference between metastatic and non-metastatic patients ($p = 0.90$, Table 1). Additionally, there is a strong correlation between the histology feature and some other features (such as CEA and nodule diameter), which could lead to multicollinearity. Furthermore, the histology feature may not always be readily available in clinical practice, making it less practical for use in a predictive model. Therefore, to enhance the stability and interpretability of the model, we chose not to include the histology feature. In the end, there were 9 features included in the next model training process (CEA, nodule diameter, CYFRA21-1, NSE, SCC, age, nodule number, nodule location, and sex, ordered by their importance).

TABLE 1 | Clinicopathological features of 286 patients in the study.

Characteristics	All, No. (%) ($n=286$)[a]	Non-metastatic ($n=148$)	Metastasis ($n=138$)	p[b]
Age (years)	62.91 (10.90)	63.26 (11.34)	62.52 (10.44)	0.48
Sex				0.23
Female	127 (44.4%)	71 (48.0%)	56 (40.6%)	
Male	159 (55.6%)	77 (52.0%)	82 (59.4%)	
Smoking				0.41
No	161 (56.3%)	87 (58.8%)	74 (53.6%)	
Yes	125 (43.7%)	61 (41.2%)	64 (46.4%)	
Exposure				0.76
No	275 (96.2%)	143 (96.6%)	132 (95.7%)	
Yes	11 (3.8%)	5 (3.4%)	6 (4.3%)	
Family tumor				1
No	253 (88%)	131 (89%)	122 (88%)	
Yes	33 (12%)	17 (11%)	16 (12%)	
Prior tumor				0.37
TOL	10 (3.5%)	5 (3.4%)	5 (3.6%)	
PTB	9 (3.1%)	7 (4.7%)	2 (1.5%)	
COPD	15 (5.2%)	6 (4.0%)	9 (6.5%)	
PF	1 (0.3%)	1 (0.7%)	0	
No	251 (87.8%)	129 (87.2%)	122 (88.4%)	
Nodule size	20.09 (6.03)	18.39 (6.20)	21.90 (5.29)	**1.21e-06**
Nodule numbers				**3.03e-05**
1	130 (45.5%)	85 (57.4%)	45 (32.6%)	
> =2	156 (54.5%)	63 (42.6%)	93 (67.4%)	
Nodule location				0.20
RUL	89 (31.1%)	43 (29.1%)	46 (33.3%)	
RML	17 (5.9%)	11 (7.4%)	6 (4.3%)	
RLL	46 (16.1%)	22 (14.9%)	24 (17.4%)	
LUL	74 (25.9%)	34 (23.0%)	40 (29.0%)	
LLL	60 (21.0%)	38 (25.7%)	22 (15.9%)	
Histology				**0.003**
OTIL	17 (5.9%)	11 (7.4%)	6 (4.3%)	
NSCLC	11 (3.8%)	2 (1.4%)	9 (6.5%)	
MIA-LUAD	2 (0.7%)	2 (1.4%)	0	
IA-LUAD	220 (76.9%)	110 (74.3%)	110 (79.7%)	
LUSC	30 (10.5%)	22 (14.9%)	8 (5.8%)	
SLC	6 (2.1%)	1 (0.7%)	5 (3.6%)	
Micropapillary				0.12
No	264 (91.6%)	133 (89.9%)	131 (94.9%)	

(Continues)

TABLE 1 | (Continued)

Characteristics	All, No. (%) ($n=286$)[a]	Non-metastatic ($n=148$)	Metastasis ($n=138$)	p[b]
Yes	22 (8.4%)	15 (10.1%)	7 (5.1%)	
CEA (ng/ml)	18.14 (61.53)	3.42 (3.14)	33.91 (85.91)	**8.58e-14**
Pro-GRP (pg/ml)	130.79 (709.72)	57.26 (27.16)	209.68 (1017.32)	0.90
NSE (ng/ml)	14.32 (21.68)	11.30 (4.52)	17.57 (30.58)	**9.39e-03**
CYFRA21-1 (ng/ml)	4.14 (6.33)	2.81 (1.98)	5.58 (8.67)	**2.41e-03**
SCC (ng/ml)	2.09 (5.89)	2.41 (5.92)	1.75 (5.85)	0.26

[a]Mean (SD); n/N (%).
[b]Wilcoxon rank sum test; Fisher's exact test. A p-value of < 0.05 was considered statistically significant.

3.3 | Model Development and Performance Evaluation

In the training dataset, with repeating 10-fold cross-validation along numerous sets of hyperparameters, we obtained the best hyperparameters for Random Forest, Gradient Boosting, and AdaBoost classifiers. We calculated the precision and AUC as classifiers performance in the validation dataset as shown in Figure 3A–C. With the best hyperparameters, the Random Forest model showed the best competence compared with the other two models. Eventually, the Random Forest model yielded an average precision of 0.93 (calculated across 10-fold cross-validation results) with an AUC of 0.92 (95% CI: 0.88–0.94) compared with the Gradient Boosting and AdaBoost classifiers in the internal validation dataset, which yielded average precision of 0.87 and 0.91 with AUCs of 0.87 (95% CI: 0.84–0.93) and 0.90 (95% CI: 0.86–0.92), respectively (calculated across 10-fold cross-validation results). Besides, the Random Forest model outperformed the other models with a higher accuracy, precision, specificity, sensitivity, and F1 score, indicating better performance. Thus, we eventually chose the Random Forest model as our prediction model for further application. Ultimately, the SHAP analysis was employed to scrutinize these three models, which demonstrated the impact of each feature on the sample and identified both positive and negative influences and emphasized the importance of certain features such as CEA and nodule diameter (Figure 3D).

3.4 | Survival Analysis

In the metastasis group, we conducted a survival analysis using nine features, which included CEA, nodule diameter, CYFRA21-1, NSE, SCC, age, nodule number, nodule location, and sex. While the five features (age, sex, nodule location, nodule number, and SCC) did not show significant differences between the subgroups, the nodule diameters, CEA, CYFRA21-1, and NSE were significantly correlated with the patient's follow-up survival. CEA and nodule diameters were the top two important features in our prediction model, and patients with higher values showed a significantly poor prognosis (Figure S2).

3.5 | Online Prediction Tool

For user-friendly access to our prediction model, we established an online platform called Metastasis of cT1 Lung Cancer Prediction (MoLPre) with nine features enrolled in our prediction model. A screenshot shows that the platform is available at https://molpre.cqmu.edu.cn/ (Figure 4). Users can predict the metastasis of lung cancer by submitting basic information into the web page of patients and pulmonary nodules. Multiple patients can also be submitted by uploading documents, except for individual patients. A sample file was provided for users, and the units of the values of the five numeric features should be consistent with the requirements in the corresponding brackets. After information submission, the results display whether the cancer of the patient tends to be metastatic.

4 | Discussion

Tumor size is regarded as an important risk factor for metastasis [4, 24]. However, some small (diameter ≤ 30 mm) lung cancers were found to have metastases at the initial diagnosis, which seriously harmed human health. Kim et al. reported that extrathoracic metastasis was apparent at the initial examination in 13% of 90 T1 lung cancer patients and at the 1-year follow-up examination in 11% of the patients [6]. Thus, predicting oncological behavior is essential when deciding between a surgical plan, aggressive surveillance, and aggressive antitumor therapy.

Within the 286 patients with cT1 solid lung cancer from the pulmonary nodule and lung cancer whole course management in our hospital, as much as possible patient information were collected in the past 4 years and a median follow-up of 17.59 [1.93–75.6] months. Nodule size, nodule numbers, histology, and tumor markers (CEA, NSE, and CYFRA21-1) were significantly associated with metastasis.

Currently, several studies have focused on predicting the metastasis in patients with lung cancer. Cho revealed that the standardized uptake values (SUV) of mediastinal lymph nodes on 18F-FDG PET/CT were predictors of nodal metastases [25]. In Andersen's study, tumor texture on CT images demonstrated a statistically significant difference between benign and metastatic lymph nodes, with a sensitivity of 53% and a specificity of 97% [26]. Kyongmin et al. developed and validated a deep cubical nodule transfer learning algorithm (DeepCUBIT), which accurately predicted lymphovascular invasion or nodal involvement in cT1 NSCLC using CT images, based on transfer learning and a 3D convolutional neural network, with a sensitivity of 0.32, specificity of 0.89, and accuracy of 0.76 [27]. However, potential

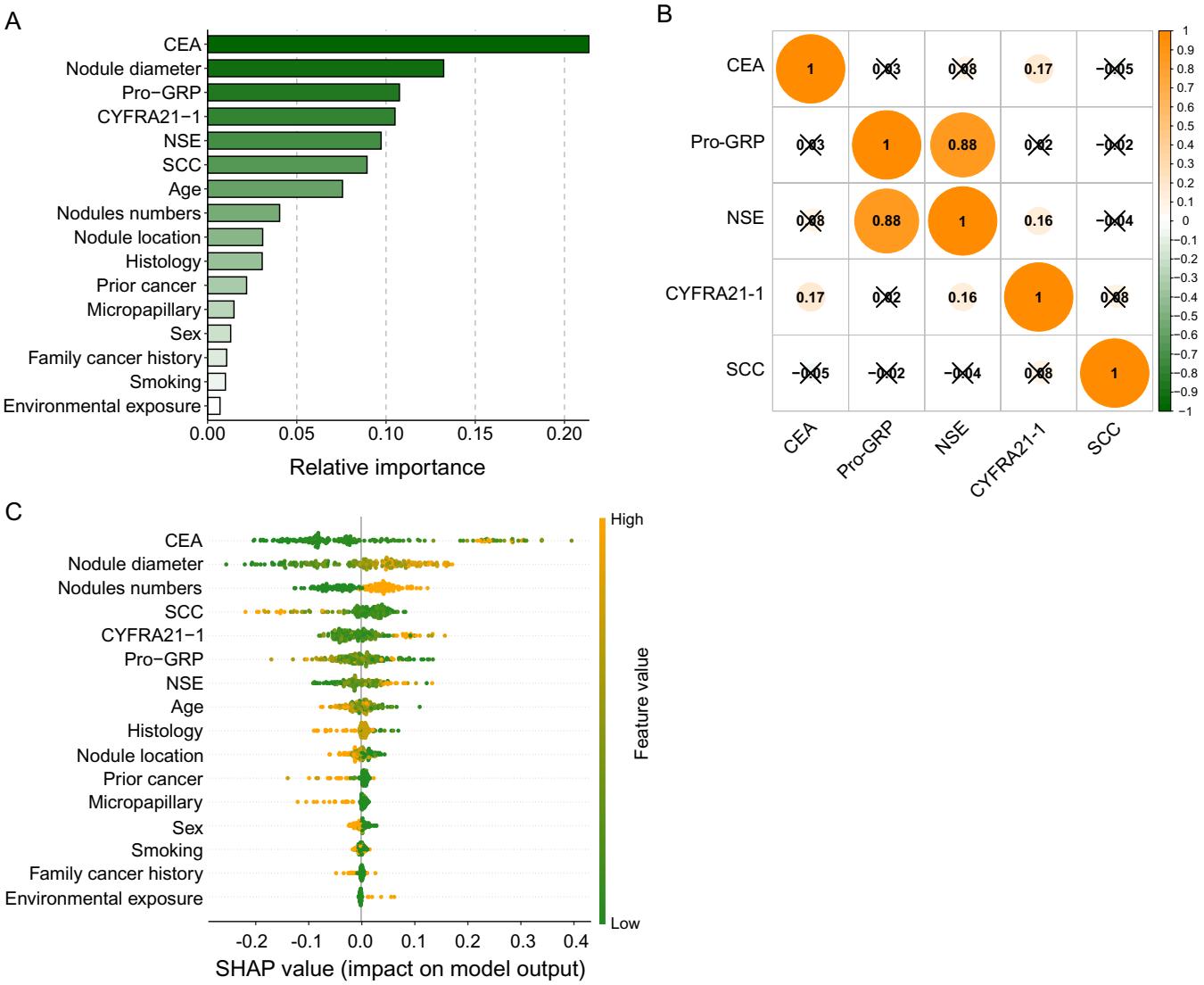

FIGURE 2 | Feature selection. (A) Bar plots show the importance of each feature measured by the Random Forest model, where the colors from white to green represent the degree from low to high importance. (B) Pair-wise correlation of 5 tumor markers. The correlation between two markers is depicted by the size, colors, and inside-numbers of each circle. The colors from green to orange indicate the correlation from low to high. There is a × if the significance of correlation has not passed the criterion ($p < 0.05$). (C) SHAP analysis was performed on Random Forest model to visually the importance of each feature.

of distant metastatic prediction in other published studies have shown generally limited success rates [28, 29]. Yi Tian developed a nomogram model to predict the occurrence of lymph node metastasis and distant metastasis in early stage NSCLC, with an AUC of 0.72 (95% CI:0.71–0.73) and AUC of 0.79 (95% CI:0.76–0.82), respectively [30]. Zhang et al. built a nomogram to predict the occurrence of brain metastases in resected NSCLC and exhibited a sufficient level of discrimination according to the C-index (0.74, 95% CI:0.67–0.82) [31]. Generally, mathematical models incorporate as many clinical features as possible, such as pathological differentiation and histological type, obtained after the invasive method. Nevertheless, no studies have been performed to predict metastasis status (lymph nodes or distal) in patients with cT1 solid lung cancer.

All of the above-mentioned published methods need to be analyzed after tumor metastasis, which is usually detected only after it has already developed to a certain stage. It remains challenging to describe the risk of metastasis of the primary tumor in advance. Our proposed machine-learning-based model, MoLPre, presented an excellent prediction ability and was characterized by its convenience for use. In the metastasis prediction process, only nine clinical factors were incorporated, which were non-invasive and convenient for clinicians to perform individualized risk prediction for each patient.

Tumor markers play a certain role in the early diagnosis and detection of metastasis in lung cancer, but their use is limited. CEA, CYFRA21-1, NSE, and SCC are commonly used tumor markers in clinical practice to assist in the diagnosis and evaluation of therapeutic effectiveness for lung cancer. However, even when these markers are detected together, they can be influenced by other factors and have limitations in diagnosing and treating lung cancer. Currently, it is believed that their ability to

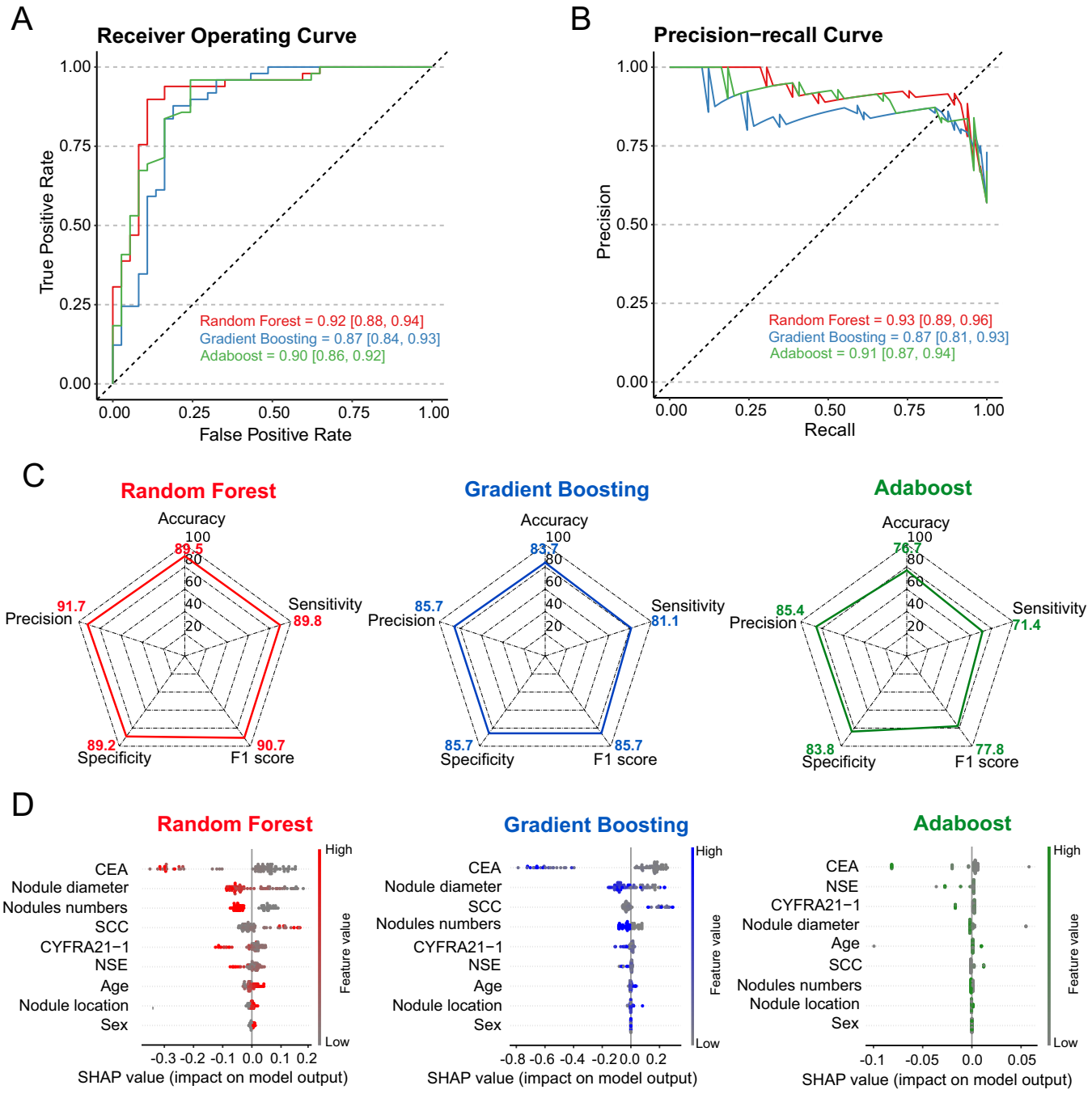

FIGURE 3 | Performance of classifiers in the validation dataset. (A) Model performance measured by AUC in the validation dataset. (B) Model performance measured by average precision (AP) in the validation dataset. (C) Radar maps of 5 diagnostic indices for three classifiers in the validation dataset. (D) SHAP analysis was performed on these models to visualize the insight on the nature of the relationship of the key predictors to the outcome.

predict early metastasis of lung cancer is inadequate. Therefore, they have little significance in guiding the diagnosis, prediction of early metastasis, and treatment of lung cancer. Factors such as nodule diameter, age, nodule number, nodule location, and sex can affect the metastasis of lung cancer. These factors are not independent; they can interact with each other and collectively affect the progression and metastasis of lung cancer. Through predictive model analysis, we can combine these meaningful factors for early metastasis of cT1-type lung cancer to identify patients who are more likely to develop metastasis in its early stages. This helps take more aggressive monitoring and treatment measures to improve patient survival rates and quality of life. Additionally, clinical predictive models can provide guidance for treatment plans. Based on individual characteristics and conditions of patients, doctors can develop personalized treatment plans. Through predictive model evaluations, doctors can understand patients' prognosis and select more suitable treatment methods. Nevertheless, it is evident that our classifier cannot achieve 100% accuracy with limited patient and information. The MoLPre web application is designed as a supportive tool for clinicians, including oncologists and pulmonologists, and scholars interested in this line of research to assist in the assessment

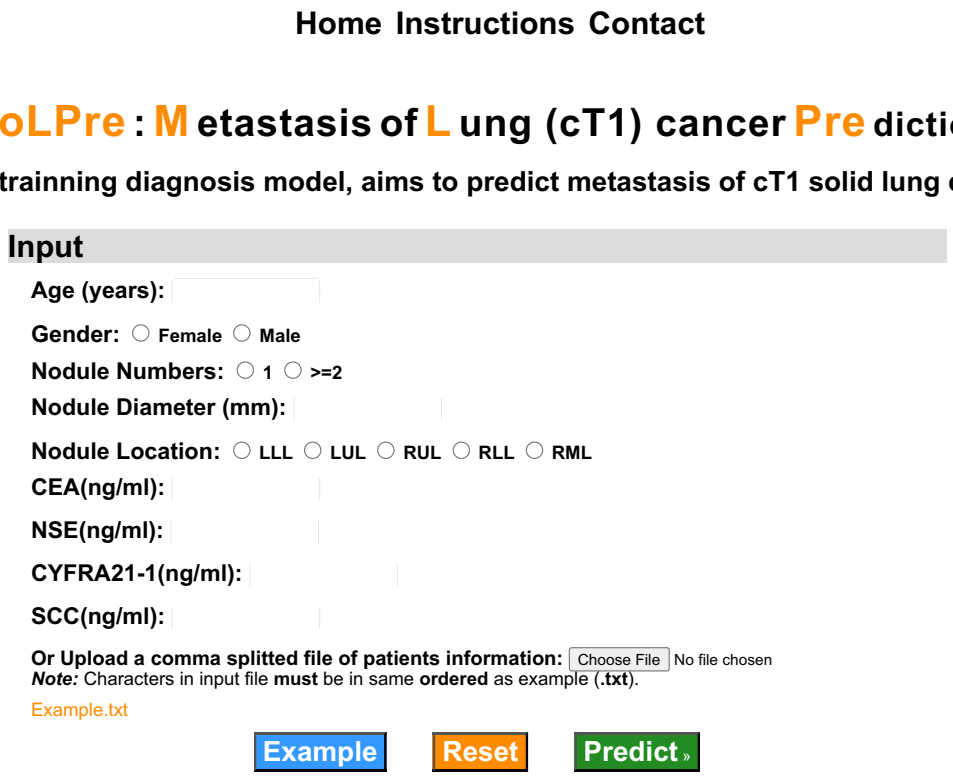

FIGURE 4 | A screenshot of the online prediction tool. The browser-based tool can be found at https://molpre.cqmu.edu.cn/ for metastasis prediction. Users can predict the metastasis of cT1 solid lung cancer by submitting nine features or multiple patients by uploading documents. MoLPre provides a sample file for reference (Example.txt, users can click it and download the example file on local computer) and the units of the value must be the same as the requirements. After submitting the information of selected patients, the results will show whether the sample is metastatic.

of metastasis risk in patients with cT1 solid lung cancer. The intended users are healthcare professionals who have access to relevant clinical data and wish to incorporate machine learning predictions into their decision-making process. This application is not a substitute for clinical judgment but serves to augment decision-making by providing an additional layer of information regarding potential metastatic risk. Users can input patient characteristics and receive predictions, which should be interpreted alongside clinical assessments and diagnostic tests.

Our study had three limitations. The first limitation was the lack of the clinical information on the patients. The SUV was confirmed to be a useful marker for predicting metastasis [25]. However, more than half of the SUV were missing, 67.7% and 65.9% in the 148 non-metastatic group and 138 metastasis group, respectively (52.6% and 61.0% of were missing in the original raw data). Therefore, we discarded these valuable markers first without any other choice. The next limitation is that our model was performed on patients collected from a single-center cohort. It is urgent to collect additional external patient data to confirm and refine our prediction tool, which is limited by time constraints and ethical issues. At present, the limitation of patients with full medical information data is an unavoidable problem. The last limitation is the lack of external validation. While we optimized the hyperparameters through cross-validation within the training dataset, this alone is not sufficient to ensure the model's generalizability. External validation using independent cohorts is essential to evaluate the model's robustness and ensure its applicability to diverse patient populations. Therefore, at this stage, the model should be viewed as preliminary and should not yet be considered for implementation in clinical practice. Future work will focus on validating the model in external cohorts from multiple centers to ensure its clinical utility.

In conclusion, metastasis prediction of cT1 solid lung cancer can help distinguish patients based on disease severity and ensure appropriate treatment. In this study, we developed a reliable prediction model with good performance for metastasis prediction and embedded this model in a web application as a user-friendly tool without registration.

Author Contributions

J.L. and H.W. wrote the manuscript. L.R. and L.Y. designed the research. J.L., J.H., W.L., and M.A. performed the research. W.Z. and J.M. analyzed the data.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers and editors, whose valuable comments greatly contributed to this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The original data used in this study will be available from the corresponding author upon request. All the source codes for our model are available at https://github.com/Jie-lan/MoLPre.

References

1. R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer Statistics, 2023," *CA: a Cancer Journal for Clinicians* 73, no. 1 (2023): 17–48, https://doi.org/10.3322/caac.21763.

2. Q. Mao, W. Xia, G. Dong, et al., "A Nomogram to Predict the Survival of Stage IIIA-N2 Non-small Cell Lung Cancer After Surgery," *Journal of Thoracic and Cardiovascular Surgery* 155, no. 4 (2018): 1784–1792, https://doi.org/10.1016/j.jtcvs.2017.11.098.

3. C. Allemani, T. Matsuda, V. Di Carlo, et al., "Global Surveillance of Trends in Cancer Survival 2000-14 (CONCORD-3): Analysis of Individual Records for 37 513 025 Patients Diagnosed With One of 18 Cancers From 322 Population-Based Registries in 71 Countries," *Lancet* 391, no. 10125 (2018): 1023–1075, https://doi.org/10.1016/s0140-6736(17)33326-3.

4. F. Bao, P. Yuan, X. Yuan, X. Lv, Z. Wang, and J. Hu, "Predictive Risk Factors for Lymph Node Metastasis in Patients With Small Size Non-small Cell Lung Cancer," *Journal of Thoracic Disease* 6, no. 12 (2014): 1697–1703, https://doi.org/10.3978/j.issn.2072-1439.2014.11.05.

5. Y. Zhang, Y. Sun, L. Shen, et al., "Predictive Factors of Lymph Node Status in Small Peripheral Non-small Cell Lung Cancers: Tumor Histology Is More Reliable," *Annals of Surgical Oncology* 20, no. 6 (2013): 1949–1954, https://doi.org/10.1245/s10434-012-2829-x.

6. K. J. Jung, K. S. Lee, H. Kim, et al., "T1 Lung Cancer on CT: Frequency of Extrathoracic Metastases," *Journal of Computer Assisted Tomography* 24, no. 5 (2000): 711–718, https://doi.org/10.1097/00004728-200009000-00008.

7. L. R. Heavey, G. M. Glazer, B. H. Gross, I. R. Francis, and M. B. Orringer, "The Role of CT in Staging Radiographic T1N0M0 Lung Cancer," *AJR. American Journal of Roentgenology* 146, no. 2 (1986): 285–290, https://doi.org/10.2214/ajr.146.2.285.

8. H. Tang, W. Liu, and K. Huang, "Stereotactic Ablative Radiotherapy for Inoperable T1-2N0M0 Small-Cell Lung Cancer," *Thorac Cancer* 13, no. 7 (2022): 1100–1101, https://doi.org/10.1111/1759-7714.14355.

9. P. F. Roberts, D. M. Follette, D. von Haag, et al., "Factors Associated With False-Positive Staging of Lung Cancer by Positron Emission Tomography," *Annals of Thoracic Surgery* 70, no. 4 (2000): 1154–1159; discussion 1159-60, https://doi.org/10.1016/s0003-4975(00)01769-0.

10. R. Kanzaki, M. Higashiyama, A. Fujiwara, et al., "Occult Mediastinal Lymph Node Metastasis in NSCLC Patients Diagnosed as Clinical N0-1 by Preoperative Integrated FDG-PET/CT and CT: Risk Factors, Pattern, and Histopathological Study," *Lung Cancer* 71, no. 3 (2011): 333–337, https://doi.org/10.1016/j.lungcan.2010.06.008.

11. D. Gompelmann, K. Kontogianni, N. Sarmand, et al., "Endobronchial Ultrasound Elastography for Differentiating Benign and Malignant Lymph Nodes," *Respiration* 99, no. 9 (2020): 779–783, https://doi.org/10.1159/000509297.

12. M. G. Dal Bello, R. A. Filiberti, A. Alama, et al., "The Role of CEA, CYFRA21-1 and NSE in Monitoring Tumor Response to Nivolumab in Advanced Non-small Cell Lung Cancer (NSCLC) Patients," *Journal of Translational Medicine* 17, no. 1 (2019): 74, https://doi.org/10.1186/s12967-019-1828-0.

13. Z. F. Jiang, M. Wang, and J. L. Xu, "Thymidine Kinase 1 Combined With CEA, CYFRA21-1 and NSE Improved Its Diagnostic Value for Lung Cancer," *Life Sciences* 194 (2018): 1–6, https://doi.org/10.1016/j.lfs.2017.12.020.

14. N. T. H. Dinh, J. Lee, J. Lee, et al., "Indoor Dust Extracellular Vesicles Promote Cancer Lung Metastasis by Inducing Tumour Necrosis Factor-α," *Journal of Extracellular Vesicles* 9, no. 1 (2020): 1766821, https://doi.org/10.1080/20013078.2020.1766821.

15. K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection," *Genome Medicine* 13, no. 1 (2021): 152, https://doi.org/10.1186/s13073-021-00968-x.

16. Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging Role of Deep Learning-Based Artificial Intelligence in Tumor Pathology," *Cancer Communications* 40, no. 4 (2020): 154–166, https://doi.org/10.1002/cac2.12012.

17. M. Avanzo, L. Wei, J. Stancanello, et al., "Machine and Deep Learning Methods for Radiomics," *Medical Physics* 47, no. 5 (2020): e185–e202, https://doi.org/10.1002/mp.13678.

18. J. M. M. Rogasch, L. Michaels, G. L. Baumgartner, et al., "A Machine Learning Tool to Improve Prediction of Mediastinal Lymph Node Metastases in Non-small Cell Lung Cancer Using Routinely Obtainable [(18)F]FDG-PET/CT Parameters," *European Journal of Nuclear Medicine and Molecular Imaging* 50, no. 7 (2023): 2140–2151, https://doi.org/10.1007/s00259-023-06145-z.

19. K. Chen, Y. Nie, S. Park, et al., "Development and Validation of Machine Learning-Based Model for the Prediction of Malignancy in Multiple Pulmonary Nodules: Analysis From Multicentric Cohorts," *Clinical Cancer Research* 27, no. 8 (2021): 2255–2265, https://doi.org/10.1158/1078-0432.CCR-20-4007.

20. Y. Zhou, J. Du, C. Ma, et al., "Mathematical Models for Intraoperative Prediction of Metastasis to Regional Lymph Nodes in Patients With Clinical Stage I Non-small Cell Lung Cancer," *Medicine (Baltimore)* 101, no. 42 (2022): e30362, https://doi.org/10.1097/MD.0000000000030362.

21. J. Wang, Y. Chu, J. Li, et al., "Development of a Prediction Model With Serum Tumor Markers to Assess Tumor Metastasis in Lung Cancer," *Cancer Medicine* 9, no. 15 (2020): 5436–5445, https://doi.org/10.1002/cam4.3184.

22. S. M. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions, in Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc, 2017), 4768–4777.

23. S. M. Lundberg, G. Erion, H. Chen, et al., "From Local Explanations to Global Understanding With Explainable AI for Trees," *Nature Machine Intelligence* 2, no. 1 (2020): 56–67, https://doi.org/10.1038/s42256-019-0138-9.

24. J. Wang, K. Welch, L. Wang, and F. M. (S.) Kong, "Negative Predictive Value of Positron Emission Tomography and Computed Tomography for Stage T1-2N0 Non-small-Cell Lung Cancer: A Meta-Analysis," *Clinical Lung Cancer* 13, no. 2 (2012): 81–89, https://doi.org/10.1016/j.cllc.2011.08.002.

25. J. Cho, J. G. Choe, K. Pahk, et al., "Ratio of Mediastinal Lymph Node SUV to Primary Tumor SUV in 18F-FDG PET/CT for Nodal Staging in Non-Small-Cell Lung Cancer," *Nuclear Medicine and Molecular Imaging* 51, no. 2 (2017): 140–146, https://doi.org/10.1007/s13139-016-0447-4.

26. M. B. Andersen, S. W. Harders, B. Ganeshan, J. Thygesen, H. H. Torp Madsen, and F. Rasmussen, "CT Texture Analysis Can Help Differentiate Between Malignant and Benign Lymph Nodes in the Mediastinum in Patients Suspected for Lung Cancer," *Acta Radiologica* 57, no. 6 (2016): 669–676, https://doi.org/10.1177/0284185115598808.

27. K. S. Beck, B. Gil, S. J. Na, et al., "DeepCUBIT: Predicting Lymphovascular Invasion or Pathological Lymph Node Involvement of Clinical T1 Stage Non-Small Cell Lung Cancer on Chest CT Scan Using Deep Cubical Nodule Transfer Learning Algorithm," *Frontiers in Oncology* 11 (2021): 661244, https://doi.org/10.3389/fonc.2021.661244.

28. H. Zhou, D. Dong, B. Chen, et al., "Diagnosis of Distant Metastasis of Lung Cancer: Based on Clinical and Radiomic Features," *Translational Oncology* 11, no. 1 (2018): 31–36, https://doi.org/10.1016/j.tranon.2017.10.010.

29. T. P. Coroller, P. Grossmann, Y. Hou, et al., "CT-Based Radiomic Signature Predicts Distant Metastasis in Lung Adenocarcinoma," *Radiotherapy and Oncology* 114, no. 3 (2015): 345–350, https://doi.org/10.1016/j.radonc.2015.02.015.

30. Y. Tian, Y. He, X. Li, and X. Liu, "Novel Nomograms to Predict Lymph Node Metastasis and Distant Metastasis in Resected Patients With Early-Stage Non-small Cell Lung Cancer," *Annals of Palliative Medicine* 10, no. 3 (2021): 2548–2566, https://doi.org/10.21037/apm-20-1756.

31. F. Zhang, W. Zheng, L. Ying, et al., "A Nomogram to Predict Brain Metastases of Resected Non-Small Cell Lung Cancer Patients," *Annals of Surgical Oncology* 23, no. 9 (2016): 3033–3039, https://doi.org/10.1245/s10434-016-5206-3.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.