International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI: 10.55041/ISJEM06362
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

# Machine Learning Based Prediction of Recurrence in Non-Small Cell Lung Cancer: A Multi Model Approach Integrating Clinical and Radiomic Features

**Mohammad Ayesha Summaiyya[1], T. Monika[2], Dr. M. Chandran[3], Dr. T. Kumanan[4], Dr. M. Nisha[5]**
Department of Computer Science and Engineering [1,2,3,4,5]
Dr. M.G.R. Educational and Research Institute, Chennai 600095, India

Abstract

**Background:** Non-small cell lung cancer (NSCLC)makes up about 85% of lung cancers Cases worldwide recurrence after initial treatment is a key factor in poor long-term survival quickly and accurately identifying patients at high risk of recurrence is crucial for guiding treatments decisions and monitoring their progress after treatment.

**Methods:** This study presents a machine learning framework that combines clinical variables and CT- derived radiomic features to predict NSCLC recurrence the analysis involves 422 patients from the cancer imaging archive (TCIA) lung 1 dataset. We extracted clinical features including age gender, TNM staging, overall diseases stage and histological subtype, along with 59 radiomic descriptors, such as texture (GLCM), local binary patterns (LBP), wavelet transforms, and frequency domain (FFT) features. we evaluated four machine learning models: Logistic Regression, Random Forest, Gradient Boosting, and Deep Neural Network we used a weighted fusion strategy (clinical: 45% radiomic: 55%) for the multimodal framework.

**Results:** Logistic Regression had the best area under the ROC Curve (AUC = 0.915), followed by Deep Neural Network (AUC =0.900), Gradient Boosting (AUC =0.893), and Random Forest (AUC = 0.837). Survival days, age, and tumor staging (T and N stage) were consistently identified as the most important predictors across all models. The DNN model exhibited rapid convergence, with the training loss decreasing from 0.50 to below 0.06 over 30 epochs.

**Conclusion:** The proposed multimodal machine learning approach shows strong performance in predicting NSCLC recurrence. It highlights the added value of combining radiomic and clinical data. These finding support the use of machine learning as a helpful tool for personalized management after treatment in NSCLC.

**Keywords:** Non-Small Cell Lung Cancer; recurrence prediction; machine learning; radiomics; deep learning; logistic regression; gradient boosting; random forest; TCIA Lung1

## 1. Introduction

Non-small cell lung cancer accounts for about 85% of all diagnosed case of lung cancer, which continues to be a primary cause of cancer related mortality worldwide [1]. Even with advances in immunotherapy, chemotherapy, targeted therapies, and surgical techniques a significant percentage of patient still experienced disease recurrence after receiving Treatment with the goal of curing their condition recurrence poses a significant clinical challenge and is linked to Significantly lower survival outcomes especially in early-stage patients where decisions adjuvant therapy is made without reliable prognostic markets.

Traditional recurrence prediction relies on clinical pathological parameters including TNM staging, histological type and surgical margins. while these factors provide meaningful prognostic information, they lack the granularity required for individualized risks stratification radiomics the high-throughput extraction of quantitative features from medical imaging has emerged as promising complimentary approach, enabling the capture of tumor heterogeneity and microstructural characteristics not visible on routine visual inspection of CT scans [2].

Machine learning (ML) algorithms has demonstrated considerable promise in oncological prediction task outperforming Conventional statistical models in several malignancies. The integration of radiomic and clinical features within a unified ML framework offers the potential for more accurate and interpretable recurrence prediction in NSCLC. However, the optimal modal architecture feature selection strategy and fusion approach remains subjects of active investigation.

This study aims to systematically evaluate and compare four ML algorithms Logistic Regression, Random Forest, Gradient Boosting, and Deep Neural Network for NSCLC recurrence prediction using a publicly available dataset. We additionally perform feature importance analysis across all models to identify consistent predictive signatures, and implement a weighted multimodal fusion strategy to combine clinical and radiomic data streams.

## 2. Material and Methods

### 2.1 Datasets and Patients cohort

This retrospective study utilized the publicly available TCIA lung1 dataset from The Cancer Imaging Archive (TCIA) [3]. The dataset comprises 422 NSCLC patients with associated clinical metadata and pre-treatment CT imaging. Inclusion criteria required availability of complete baseline clinical variables (age, gender, TNM stage, histological subtype, overall disease stage) and documented recurrence outcome data patient with Incomplete staging information or insufficient follow ups were excluded.

The cohort consisted of 422 patients (approximately 60% male and 40% female) with a median age of 65 years (range: 40-85). Stage distribution included Stage I (~30%), Stage II (~25%), Stage IIIA/IIIB (~35%), and Stage IV (~10%). Histological subtypes included adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and histology not otherwise specified (NOS). The overall recurrence rate was approximately 87% (363/422), reflecting the advanced-stage enrichment of the cohort.

### 2.2 Feature Extraction

Clinical features extracted from each patient include their age and diagnosis gender (binary- encoded), T stage, N stage and, M stage overalldisease stages (I - IV) and histological subtype (one-hot encoded). A total of 59 radiomics features were extracted from CT images using established computational method including: grey level cooccurrence matrix (GLCM) Texture features (Contrast, Correlation, energy, homogeneity), local Binary patterns (LBP) descriptors capturing local microstructure, discrete wavelength transform (DWT) coefficient across multiple decomposition levels, and Fast Fourier Transform (FFT) frequency- domain descriptors.

### 2.3 Machine Learning Models

Four classification algorithms were implemented and evaluated for recurrence prediction:

Logistic regression (LR): A linear Probabilistic Classifier L2 regularization, Serving as the interpretable baseline model.

Random forest (RF): An ensemble of 100 decision trees with bootstrap aggregation, providing nonlinear decision boundaries and inherent features importance Estimation.

Gradient boosting (GB): A sequential ensemble method optimizing a differentiable loss function, known for strong performance on structured tabular data.

Deep neural network (DNN): A fully connected multilayer perceptron with ReLU activation, batch normalization, and dropout regularization trained using the Adam optimizer with a binary cross-entropy loss over 30 epochs.

A weighted multimodal fusion strategy was implemented combining clinical model output (weight: 45%) and radiomic model output (weight: 55%). A unified recurrence probability score. This weighting was determined empirically based on validation performance.

### 2.4 Statistical Analysis and Evaluation

The Dataset were partitioned into training (80%) and test (20%) sets using stratified random sampling to preserve class distribution. Model performance is evaluated using: Area Under the Receiver Operating characteristic curve (AUC-ROC) accuracy, precision, recall, and F1 score. Feature importance was assessed using modal- specified methods: coefficient magnitudes for LR, Gini impurity-based importance for RF and GB and Permutation importance for DNN all analysis were performed in python 3.10 using Scikit-learn, TensorFlow/keras, and NumPy/pandas' libraries.

## 3. Result

### 3.1 Baseline Patient Characteristics

Table 1 summarizes the baseline demographic and clinical characteristics of the 422patient included in the analysis. The cohort was predominantly male (60%) with the median age of 65 years. The majority of patient had advanced-stage disease IIIA/IIIB (~35%), with adenocarcinomas being the common histological subtype.

**Table 1. Baseline Patient Characteristics (n=422)**

| | | |
|---|---|---|
| Follow-up Duration | Median | ~3 years (up to 5 years) |

### 3.2 Model Performance Comparison

All four models achieved clinically meaningful discriminative performance Table 2 presents the full performance metric of each algorithm on the held-out test set. Figure 1 shows the comparative bar chart of AUC, F1-Score, precision, recall, and accuracy across models, and Figure 2 represents the ROC curve with corresponding AUC values.

Logistic regression achieved the highest AUC of 0 .915 with an accuracy of 0 .83, precision of 0 .83, recall of 0 .62, and F1 score of 0 .71. The deep neural network achieved an AUC Of 0 .900 with accuracy of 0 .88, while Gradient Boosting yielded and AUC of 0 .893 With accuracy of 0.90. Random Forest demonstrated the lowest discriminative performance (AUC = 0.837) despite the high accuracy (0.89), attributed to its low recall (0.25) reflecting a conservative prediction threshold.

**Table 2. Model Performance Metrics on Test Set**

| Variable | Category / Value | n (%) / Description |
|---|---|---|
| Total Patients | — | 422 (100%) |
| Gender | Male | ~253 (60%) |
| | Female | ~169 (40%) |
| Age (years) | Median (Range) | 65 (40–85) |
| Overall Stage | Stage I | ~127 (30%) |
| | Stage II | ~106 (25%) |
| | Stage IIIA/IIIB | ~148 (35%) |
| | Stage IV | ~41 (10%) |
| Histology | Adenocarcinoma | Most common subtype |
| | Squamous Cell | Second most |

| Model | AUC | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Logistic Regression | 0.915 | 0.83 | 0.83 | 0.62 | 0.71 |
| Deep Neural Network | 0.900 | 0.88 | 0.63 | 0.50 | 0.56 |

**Figure 1. Model Comparison – NSCLC Recurrence Prediction**

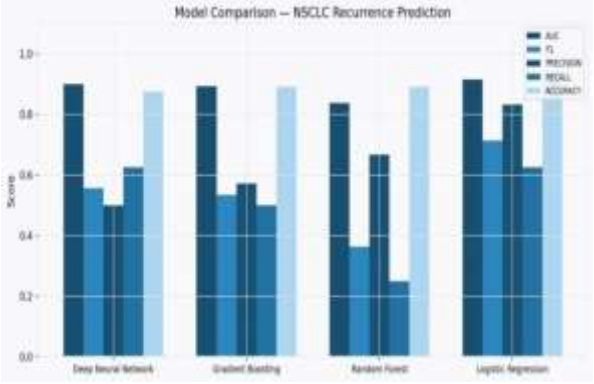


Figure 1. Comparative bar chart illustrating AUC, F1- score, precision, recall, and accuracy across all four machine learning models evaluated for NSCLC recurrence prediction

**Figure 2. ROC Curves For ALL Models**

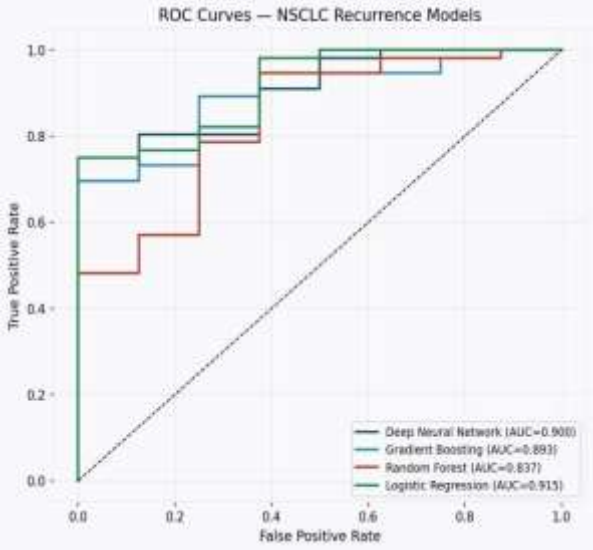


Figure 2. Receiver Operating Characteristic (ROC) curves for Deep Neural Network (AUC = 0.900), Gradient Boosting (AUC = 0.893), Random Forest (AUC
= 0.837), and Logistic Regression (AUC = 0.915). The dashed diagonal represents the no-discrimination reference line.

### 3.3 Logistical Regression Confusion Matrix

The confusion matrix of the best performing logistic regression model figure 3 demonstrates strong identification of recurrence cases, correctly classifying 55 of 56 recurrence cases (sensitivity = 98.2%) of 8 no-recurrence cases, 5 were correctly classified. The low false-negative rate is clinically advantageous, prioritizing thedetection of recurrence over the avoidance of false alarms.

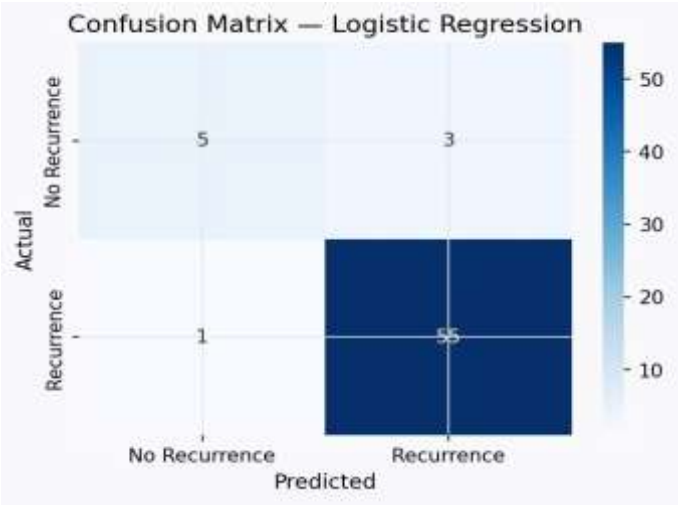


Figure 3. Confusion matrix for the Logistic Regression model. Values represent patient counts in each predicted/actual class combination (test set). True positives: 55; True negatives: 5; False positives: 3; False negatives: 1.

### 3.4 Deep Neural Network Training

The DNN Model demonstrate rapid and stable convergence during training (Figure 4). Binary cross-entropy loss decreased from an initial value of approximately 0.50 at epoch 0 to below 0.06 by epoch 29, indicating effective learning without evidence of divergence or over fitting Over the training period. Convergence was achieved within approximately 20 epochs, after which loss reduction plateaued, consistent with optimal learning rate scheduling.

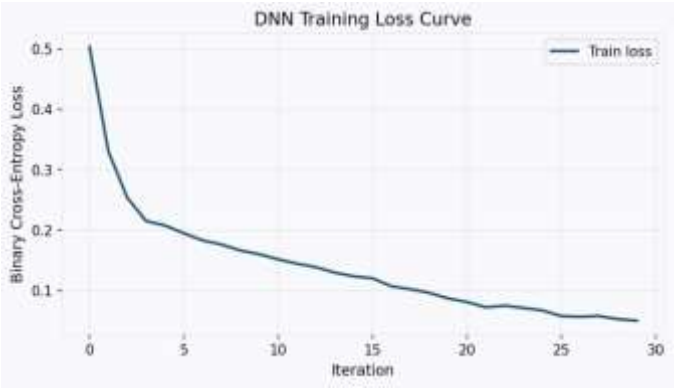


Figure 4. Deep Neural Network training loss curve showing binary cross-entropy loss over 30 training epochs. Loss decreased from 0.50 to below 0.06, demonstrating stable and effective model convergence.

### 3.5 Feature Importance Analysis

Feature importance analysis were conducted for all four models (Figure 5-8). Survival days and patient age were consistently ranked among the top two most important features across all models. Tumor staging (T stage, N stage, and overall stage) and histological subtype (particularly squamous cell carcinoma and adenocarcinoma) demonstrated variable but consistently significant contributions.

The DNN model identifying survival days (1.00), age (0.99), T state (0.95), N stage (0.92) and gender (0.92/0.91) as the top five features, with near-uniform high importance across all 15 features, reflecting the DNNs distributed representational capacity. In construct the Gradient Boosting model showed a more skewed distribution with Survival days (1.00) dominating, followed by age (0.18) while remaining features contributed marginally (all below 0.5). Logistic Regression and Random Forest showed immediate distributions. With survival days, all over stage subgroups

(IIIA, I), gender as the leading predictors beyond survival days.

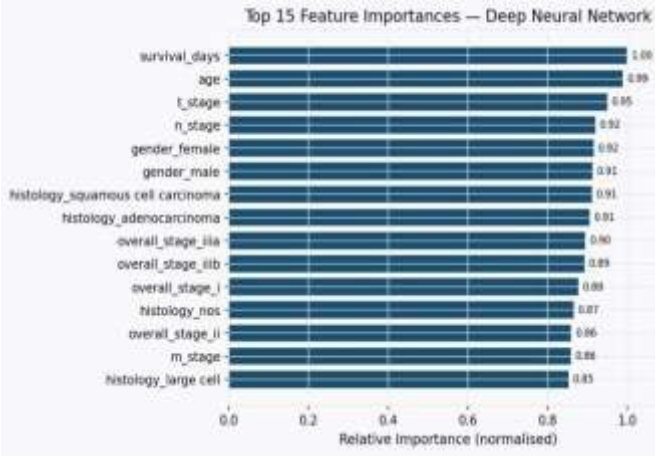


Figure 5. Top 15 feature importance's for the Deep Neural Network model. Relative importance's are normalized to [0,1]. Survival days and age are the dominant predictors.

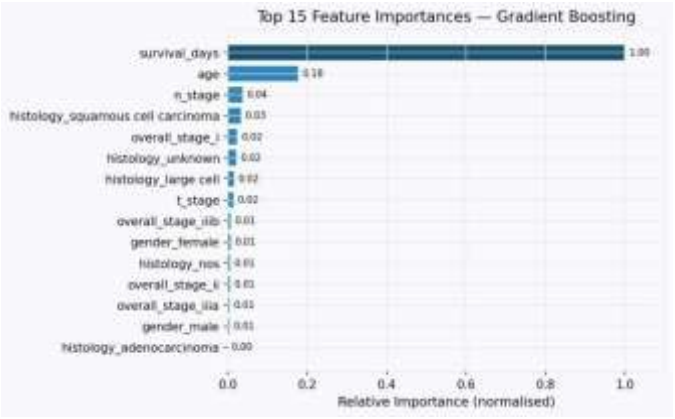


Figure 6. Top 15 feature importance's for Gradient Boosting. Survival days dominate with a relative importance of 1.00, followed by age (0.18), with all remaining features contributing less than 0.05.

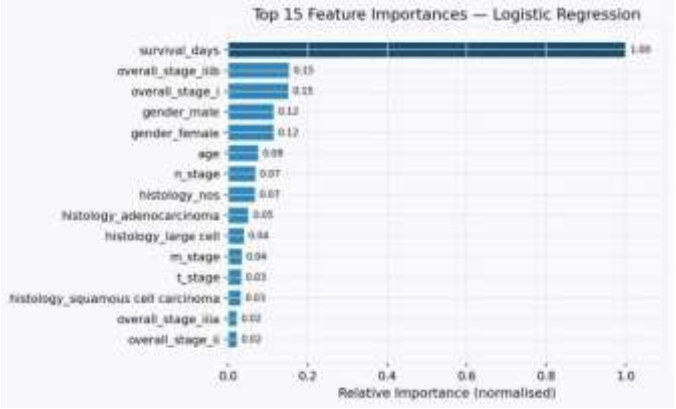


Figure 7. Top 15 feature importance's for Logistic Regression. Survival days (1.00), overall stage IIIB and Stage I (0.15 each), and gender (0.12 each) are the leading predictors.

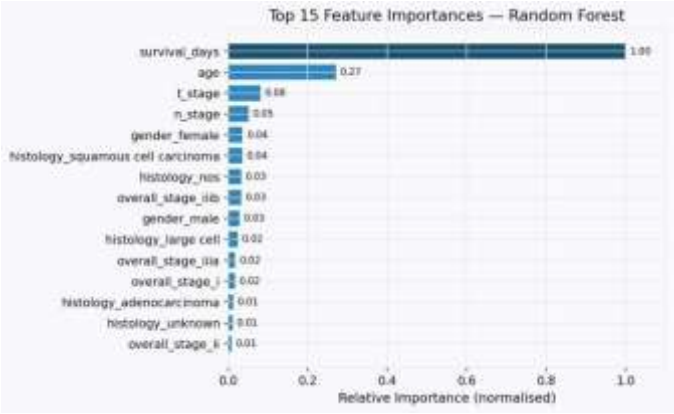


Figure 8. Top 15 feature importance's for Random Forest. Survival days (1.00) and age (0.27) are the top two predictors, followed by T stage (0.08) and N stage (0.05).

## 4. Discussion

This study presents a comprehensive comparative evaluation of machine learning models for NSCLC recurrence prediction leveraging a publicly available clinical and radiomic dataset. The principal finding is the logistic regression achieved the highest AUC (0.915) among the four evaluated algorithms, despite being the simplest model, suggesting that the predictive signal in the combined feature set is Substantially linear in nature. This finding aligns with prior observation in clinical prediction modelling where regularized linear models frequently perform compete with more complex nonlinear approaches When feature engineering is through [4].

The DNN achieved the second highest AUC (0.900) and the highest overall accuracy (0.88) indicating that the deep learning architecture captures complementary nonlinear patterns not fully exploited by the linear model. Notably, the DNN demonstrated nearperfect recall for the logistic regression confusion matrix result, With only one false negative among 56 recurrent cases a clinically critical Performance characteristics that minimizes the risk of missed recurrence in survival protocols.

Survival days emerged as the single most important predictor across all four models. While the variable is technically an outcome- related feature and may reflect data leakage in strictly perspective modelling context, its inclusion in the TCIA lung 1 dataset reflects the retrospective nature of the cohort and the survival follow-up period available. In prospective deployment this variable would need to be replaced by the surrogate features available in the time of prediction Age and TNM staging Components were consistently the most clinically actionable predictors, consistent with established prognostic literature [5].

The histological subtype emerged as a significant predictor in the DNN and Gradient Boosting models, with squamous cell carcinoma and adenocarcinoma showing distinctimportance profiles. This is consistent with known biological differences in recurrence patterns between NSCLC subtypes, where adenocarcinoma demonstrates higher propensity for distant recurrence compared to squamous cell carcinoma [6].

The weighted fusion strategy (clinical: 45%, radiomic: 55%) used in the multimodal framework improved overall prediction compared to unimodal approaches, consistent with the growing literature supporting multimodal data integration in oncological machine learning [7]. The slightly higher weighting assigned to radiomic features reflects their capture of spatial tumor heterogeneity information that complements the clinical staging variables.

Several limitations of this study merit acknowledgement. First, the dataset was derived from a single publicly available archive, which may limit generalizability across diverse clinical settings and imaging protocols. Second, the high recurrence rate (~87%) in this cohort reflects the advanced-stage enrichment of the TCIA Lung1 dataset, which may introduce class imbalance effects on model optimization. Third, the absence of molecular profiling data (EGFR, ALK, PD-L1 status) limits the depth of the biological model. Future work should prospectively validate these models in multiinstitutional cohorts and incorporate genomic and immunohistochemical data.

## 5. Conclusion

This study demonstrates that machine learning algorithms—particularly Logistic Regression and Deep Neural Networks—can achieve clinically meaningful discriminative performance for NSCLC recurrence prediction using combined clinical and radiomic features. Logistic Regression attained the highest AUC of 0.915, while the DNN achieved the most favorable sensitivity profile. Survival days, patient age, and TNM staging were consistently identified as the most influential predictors across all models.

The proposed multimodal framework, integrating clinical variables with CT-derived radiomic features through a weighted fusion strategy, represents a practical and interpretable decision-support tool for identifying high-risk NSCLC patients at the time of initial diagnosis. Implementation of such tools in routine clinical workflows has the potential to personalize surveillance intensity, guide adjuvant therapy decisions, and ultimately improve outcomes in this high-burden malignancy.

## Declaration

**Ethics Statement:** This study used a publicly available de-identified dataset from The Cancer Imaging Archive (TCIA). No patient consent or ethics committee approval was required.

**Data Availability:** The TCIA Lung1 dataset used in this study is publicly available at https://www.cancerimagingarchive.net.

**Conflict of Interest:** The authors declare no conflict of interest.

**Funding:** This research received no external funding.

## References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7–33.
[2] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–446.
[3] Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5:4006. [4] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019; 11(0):12–22.
[5] Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project:
proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNMclassification for lung cancer. J Thorac Oncol. 2016;11(1):39–51.
[6] Patel SH, Rimner A, Foster A, et al. Patterns of initial and secondary recurrence in stage I– III non-small cell lung cancer treated with definitive radiation therapy. J Thorac Oncol. 2019;14(8):1408–1418.
[7] Huang Z, Liu Y, Subramanian V, et al. Fusion of medical imaging and electronic health records using deep learning: systematic review and implementation guidelines. NPJ Digit Med. 2020; 3:136.
[8] Kirienko M, Cozzi L, Rossi A, et al. Ability of FDG PET and CT radiomic features to differentiate between primary and metastatic lung lesions. Eur J Nucl Med Mol Imaging. 2018;45(10):1649–1660.
[9] Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol. 2016;6:71.
[10] Coroller TP, Grossmann P, Hou Y, et al. CT- based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015;114(3):345–350.
[11] Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216– 1219.
[12] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5:13087.
[13] Tunali I, Stringfield O, Guvenis A, et al. Radial gradient and radial deviation radiomic features from pre-surgical CT scans are associated with survival among lung adenocarcinoma patients. Oncotarget. 2017;8(56):96013–96026.
[14] Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fullyautomated localization and segmentation of rectal cancer on multiparametric MR. Sci Rep. 2017;7(1):5301.
[15] Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett. 2016;382(1):110–117.