

Prediction of Distant Metastasis in Renal Cell Carcinoma Using Machine Learning Algorithms: A Multicenter Cohort Study

Cancer Control
Volume 33: 1–12
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10732748261432272
journals.sagepub.com/home/ccx


Yajian Li^{1,2,*}, Xinwei Wang^{1,2,*}, Moxuan Wang³, Jianzhong Shou^{1,2}, Cancan Chen⁴, and Li Wen^{1,2} 

Abstract

Introduction: Few machine learning (ML) studies have investigated the prediction of distant metastasis in patients with renal cell carcinoma (RCC). This retrospective study aimed to develop and validate predictive models based on ML algorithms for RCC patients with distant metastasis. **Methods:** We extracted RCC data from the SEER database between 2004 and 2015 (n=192,912) and from the Chinese National Cancer Center (CNCC) database between 2010 and 2020 (n=3,034). Seven different algorithms were applied to predict distant metastasis in RCC. Fivefold cross-validation was employed for model construction. The data were analyzed by using Python based on incomplete data, complete data, upsampling data and downsampling data. **Results:** After data cleaning and screening, 121,741 cases from the SEER dataset and 2803 cases from the CNCC external test set were retained. For incomplete data, the neural network model [area under the curve (AUC) 95% confidence interval (CI) of the external data: 0.7467±0.0573] achieved the highest accuracy. For the complete data, the support vector machine (SVM) model achieved the highest accuracy, with an AUC 95% CI of 0.8221±0.0485. The disparity between positive and negative samples significantly varied across the different datasets. Upsampling and downsampling analyses were also conducted. For the upsampling data, the extreme gradient boosting (XGBoost) model demonstrated the highest accuracy, with an AUC 95% CI for the external data of 0.8162±0.0558. For the downsampling data, the SVM model achieved the highest accuracy, with an AUC 95% CI of 0.8274±0.0546 for the external data. **Conclusions:** Our study revealed that ML algorithms can effectively predict distant metastasis in patients with RCC. ML models exhibit favorable application prospects in clinical practice.

Keywords

machine learning, renal cell carcinoma, prediction, distant metastasis, model

Received: 8 November 2025; Revised: 2 February 2026; Accepted: 17 February 2026

¹Department of Urology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

²Beijing Key Laboratory of Urologic Cancer Cell and Gene Therapy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

³Qingdao Economic and Technological Development Zone No. 4 Junior Middle School, Qingdao, China

⁴School of Computer Engineering, Jiangsu Ocean University, Lianyungang, China

*The authors contributed equally to this paper

Corresponding authors:

Cancan Chen, School of Computer Engineering, Jiangsu Ocean University, No. 59 Cangwu Road, Lianyungang City 222005, PR China.

Email: chencc@jou.edu.cn

Li Wen, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Panjiayuan Nanli 17#, Beijing 100021, PR China.

Email: wenli_cicams@163.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Introduction

Renal cell carcinoma (RCC) is among the most common malignant tumors affecting the urinary system. The incidence of renal cancer in China has demonstrated a continuous upward trend in recent years, with approximately 73,700 new cases being reported in 2022 and a standardized incidence rate of approximately 6.7 per 100,000 being observed.¹ RCC primarily comprises three histopathological types: clear cell RCC, papillary RCC, and chromophobe RCC, with clear cell RCC accounting for 70–80% of cases, papillary RCC accounting for 10–15%, and chromophobe RCC accounting for 5%.² The survival of RCC patients has improved due to surgical management and the use of targeted and immune drugs.^{3,4} However, a substantial proportion of RCC patients present with high-stage disease, with a poor prognosis being determined because of distant metastases.^{5–7} Thus, tumor metastasis is a critical factor in determining the prognosis of RCC patients.

Seventy percent of patients are diagnosed with stage I RCC, and 11% of patients are diagnosed with stage IV RCC.⁸ The lungs are the most commonly affected site of metastasis, followed by regional lymph nodes, the brain, bone, and soft tissue. Lee CH et al⁹ analyzed the Korean Renal Cancer Study Group metastatic RCC (mRCC) database. The results revealed the most common sites of metastasis (>5%), and the median cancer-specific survival (CSS) ranged from 13.9 (liver) to 29.1 months (lungs). Through continuous efforts in RCC research, for advanced RCC or mRCC, combinations of immune checkpoint inhibitors or combinations of immune checkpoint inhibitors with tyrosine kinase inhibitors are associated with a tumor response of 42% to 71%, with a median overall survival of 46 to 56 months being observed.⁸ These results indicate that the prognosis of mRCC patients is still poor, regardless of the presence of lymph node metastasis or other organ metastasis.^{9,10} The early detection of distant metastasis is the top priority of clinical work because the appropriate treatment options (such as surgery, immunotherapy, and targeted therapy) can be chosen correctly.

The identification of the clinicopathological risk factors that promote distant RCC metastasis is vital. In recent years, several clinical diagnostic tools have been built using different prognostic and prediction models. Although predictive factors and models for mRCC have been developed, the limited data commonly restrict individualized accurate prediction. These studies exhibit significant limitations, such as the use of older tools and limited data. The Surveillance, Epidemiology, and End Results (SEER) database represents 28% of the current U.S. population; moreover, it includes large, multi-institutional patients and could be an essential source of mRCC research to provide greater statistical power. Additionally, machine learning (ML), which is a major subbranch of artificial intelligence, is a promising statistical method that can handle large amounts of heterogeneous data. ML techniques have not been used in many studies, and some models lack interpretability to fully capture the complex relationships among variables and to provide clinically actionable explanations. Previously, we used ML algorithms as auxiliary tools to predict overall survival in RCC patients.¹¹ This technique could be used to quantify the possibility of recurrence in patients and to help with more individualized postoperative clinical management.

In our study, we aimed to develop and validate an explainable ML model for the prediction of RCC metastasis. We collected a large amount of data from the SEER database and the Chinese National Cancer Center (CNCC) dataset. The RCC cases were subjected to data standardization and included multiple clinical parameters, such as tumor grade, side, pathological type, T stage, and N stage. The RCC M0/M1 metastasis prediction algorithm was explored and visualized.

Materials and Methods

Data Collection

The data used in this study were consecutively collected from the SEER database by using SEER*Stat software (version 8.3.6) and the CNCC. A data agreement form was signed and submitted to the SEER administration. This trial was a retrospective study and was approved by the Institutional Review Committee of the CNCC (Institutional Review Board number: 21/405-3076; date: October 13, 2021). The requirement for obtaining informed consent was also waived by the Institutional Review Committee. The study was conducted in accordance with the Helsinki Declaration of 1975, as revised in 2024. All of the patient details have been de-identified. The reporting of this study conforms to the TRIPOD + AI statement.¹²

The variables included race, sex, age, marital status, tumor location, tumor size, histological type, tumor grade, tumor-node-metastasis (TNM) stage information, and surgical treatment. The SEER dataset was utilized as the internal dataset for model construction and performance evaluation, whereas the dataset from the CNCC served as an external independent dataset to test the generalizability of the models.

Data Preparation

We extracted RCC data, including data from the SEER database between 2004 and 2015 (n=192,912) and from the CNCC database between 2010 and 2020 (n=3,034). After data cleaning and screening, 121,741 cases remained in the SEER

dataset, and 2803 cases remained in the CNCC external test set. The flowchart is shown in [Figure 1](#). Given the substantial volume of the SEER data and the presence of incomplete data for some features, the dataset was initially divided into training and testing sets at a 9:1 ratio, thereby resulting in approximately 109,567 cases for the training set and 12,174 cases for the testing set. From all of the cases with complete data, 12,174 cases were randomly selected as the testing set, with the remaining 109,567 cases being allocated to the training set. Among the 109,567 cases, 80,119 cases had complete data, and 29,448 cases had incomplete data.

Data Preprocessing

To facilitate extraction and analysis, numerical characterization of the clinicopathological data of the RCC patients was initially conducted through different numbers of assignments.

Processing of Incomplete Data

The random forest interpolation method was used to supplement missing values. First, a random forest regression model was trained based on available complete data. The average values were subsequently used to supplement all of the missing features. Finally, the random forest regression model was used to regress and predict each missing feature one by one, along with replacing the missing values with the original average values.

Feature Discretization

To improve the generalizability of the model, continuous features such as age and tumor size were discretized and binarized. This study employed a feature discretization method grounded in information entropy, which is a supervised approach. Information entropy serves as a metric for quantifying the uncertainty of an event, and higher entropy values indicate greater uncertainty. The following formula was utilized for calculating information entropy:

$$IE(X) = - \sum_{i=1}^k p(x_i) \log p(x_i)$$

where X represents all of the possible categories of the sample, k denotes the total number of categories, and $p(x_i)$ signifies the probability of the sample belonging to category i . The discretization method identifies the threshold that maximizes the decrease in the entropy of category information in the training set. Based on this threshold, all of the features are discretized into binary values.

ML models were constructed by using fivefold cross-validation based on the training set of 80,119 cases with complete data and the supplemented training set of 109,567 cases via a random forest regressor. Additionally, the internal test set included 12,174 cases, and the external test set included 2,803 cases. All of the cross-validation splits were stratified to maintain the same proportion of metastatic and nonmetastatic cases in each fold as in the overall development set. The specific division results are shown in [Tables 1](#) and [2](#).

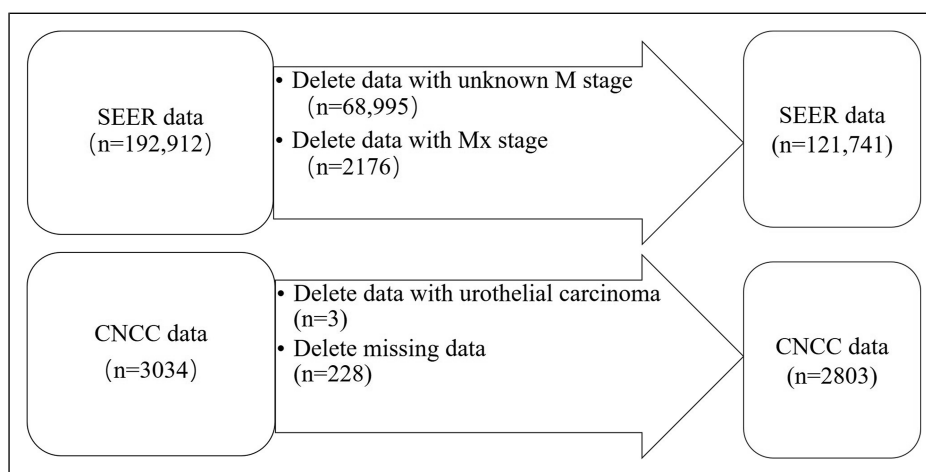


Figure 1. Flowchart for SEER and CNCC data screening

Table 1. 80,119 Cases With Complete Training Set

Set	M0 num		M1 num		Total
	Train	Validation	Train	Validation	
Fold1	59363	14827	4732	1197	80119
Fold2	59347	14843	4748	1181	80119
Fold3	59391	14799	4704	1225	80119
Fold4	59305	14885	4790	1139	80119
Fold5	59354	14836	4742	1187	80119
Test data	11296		878		12174
External data	2718		85		2803

Stability Assessment

The “leave-top1-out” approach was used to evaluate feature ranking stability: (i) selected the top-n features from the full dataset (Set1); (ii) removed the highest-ranked feature and retrained the model on the reduced dataset; (iii) extracted the new top-(n–1) features (Set2); (iv) systematically compared the ranking orders between Set 1 and Set 2 using Spearman’s rank correlation coefficient and the Jaccard similarity index.

Statistical Analysis

We trained all of the ML models by using Python version 3.7.1, NumPy version 1.20.1, Scikit-Learn version 1.0.2, SciPy version 1.7.3, and XGBoost version 1.6.2. The ML models were developed and evaluated by using 5-fold cross-validation. The receiver operating characteristic (ROC) curve, decision curve analysis (DCA), calibration curve, Delong test and SHapley Additive exPlanations (SHAP) images were used to evaluate the accuracy of the model. Both F1(1) and F1(0) scores were calculated to provide a comprehensive performance evaluation.

Results

Clinical Characteristics of the Patients

A total of 192,912 cases from the SEER dataset and 3,034 cases from the CNCC dataset were included in this study. The clinicopathological features of RCC patients from the SEER database have been described in our previous study.¹¹ The median age of the patients in the CNCC database was 58 years [interquartile range (IQR), 51–65 years]. There were 2,056 (67.8%) males and 978 (32.2%) females included in the study. Most of the RCC patients were married, accounting for 96.6% (n=2,929) of the cases. The population mostly consisted of Asian individuals (99.7%, n=3,028). The mean tumor size was 4.2±1.7 cm. At the initial diagnosis, 1.8% (n=53) and 3.0% (n=92) of the patients demonstrated lymph node and organ metastases, respectively. Histologically, the common pathological types included clear cell RCC (87.1%, n=2,642), papillary RCC (3.3%, n=101), chromophobe RCC (4.2%, n=127) and others (1.1%, n=32). The clinicopathological features of the RCC patients from the CNCC cohort are shown in [Table 3](#).

Table 2. 109,567 Cases With Incomplete Data in Training Set

Set	M0 num		M1 num		Total
	Train	Validation	Train	Validation	
Fold1	76969	19190	10684	2724	109567
Fold2	76897	19262	10756	2652	109567
Fold3	76924	19235	10730	2678	109567
Fold4	76911	19248	10743	2665	109567
Fold5	76935	19224	10719	2689	109567
Test data	11296		878		12174
External data	2718		85		2803

Table 3. Patient Characteristics at Baseline of the Chinese National Cancer Center Dataset

Characteristics	n (%)
Sex	
Male	2,056 (67.8)
Female	978 (32.2)
Median age (interquartile range, year)	58 (51-65)
Marital status	
Married	2,929 (96.6)
Unmarried	71 (2.3)
Unknown	34 (1.1)
Race	
Yellow	3,028 (99.7)
Black	2 (0.1)
White	2 (0.1)
Unknown	2 (0.1)
Tumor location	
Left	1,491 (49.1)
Right	1,535 (50.6)
Bilateral	8 (0.3)
Tumor size (mean $\pm$ SD, cm)	4.2 $\pm$ 1.7
Histological types	
Clear cell	2,642 (87.1)
Papillary	101 (3.3)
Chromophobe	127 (4.2)
Others	32 (1.1)
Unknown	132 (4.3)
Tumor grade	
G1/2	2494 (82.2)
G3/4	310 (10.2)
Unknown	230 (7.6)
T stage	
T1/2	2649 (87.3)
T3/4	385 (12.7)
N stage	
N0	2,968 (97.8)
N1	53 (1.8)
Unknown	13 (0.4)
M stage	
M0	2,942 (97.0)
M1	92 (3.0)

Model Construction

According to the aforementioned data partitioning table, it was necessary to investigate and compare the performance of the models that were constructed with either complete data or incomplete data. Additionally, as shown in [Tables 1](#) and [2](#), there was a significant disparity observed between the numbers of positive and negative samples within each dataset. Consequently, the performance of the models was compared by using upsampling and downsampling methods. The research framework was divided into the following two sections.

Incomplete Data (n=109,567) vs. Complete Data (n=80,119)

We used a training set with incomplete data, which underwent missing value processing. Fivefold cross-validation was employed for model construction. The integrated model prediction results are shown in [Table 4](#). According to the results of the test and external datasets, the SVM [area under the curve (AUC) 95% confidence interval (CI) of the test data: 0.8328 ± 0.0165 ; AUC 95% CI of the external data: 0.5874 ± 0.0884] was not effective for predicting the metastasis of RCC patients. The Bayes (AUC 95% CI of the test data: 0.869 ± 0.0123 ; AUC 95% CI of the external data: 0.7399 ± 0.0623), decision tree (AUC 95% CI of the test data: 0.8639 ± 0.0132 ; AUC 95% CI of the external data: 0.7398 ± 0.0593), logistic regression (AUC 95% CI of the test data: 0.8755 ± 0.0126 ; AUC 95% CI of the external data: 0.739 ± 0.068), neural network (AUC 95% CI of the test data: 0.8655 ± 0.0129 ; AUC 95% CI of the external data: 0.7467 ± 0.0573), random forest (AUC

Table 4. The Integrated Model Prediction Results After Five-Fold Cross-Validation in the Incomplete Data

Model	Set	Auc.	Acc.	Sens.	Spec.
Bayes	Train data	0.848±0.004	0.766±0.003	0.824±0.007	0.758±0.003
	Test data	0.869±0.012	0.821±0.007	0.794±0.027	0.823±0.007
	External data	0.740±0.062	0.823±0.014	0.667±0.103	0.828±0.014
Decision Tree	Train data	0.881±0.003	0.861±0.002	0.745±0.008	0.878±0.002
	Test data	0.864±0.013	0.839±0.006	0.789±0.027	0.843±0.007
	External data	0.740±0.059	0.755±0.015	0.690±0.101	0.757±0.016
Logistic	Train data	0.873±0.003	0.848±0.002	0.755±0.007	0.861±0.002
	Test data	0.876±0.013	0.823±0.007	0.797±0.027	0.826±0.007
	External data	0.739±0.068	0.846±0.013	0.655±0.106	0.852±0.013
Neutral network	Train data	0.881±0.003	0.861±0.002	0.745±0.008	0.878±0.002
	Test data	0.866±0.013	0.842±0.006	0.782±0.027	0.847±0.007
	External data	0.747±0.057	0.758±0.015	0.690±0.101	0.760±0.015
Random Forest	Train data	0.881±0.003	0.861±0.002	0.745±0.008	0.877±0.002
	Test data	0.864±0.013	0.834±0.006	0.797±0.027	0.837±0.007
	External data	0.743±0.060	0.759±0.015	0.690±0.101	0.761±0.016
SVM	Train data	0.834±0.004	0.847±0.002	0.760±0.008	0.859±0.002
	Test data	0.833±0.017	0.847±0.006	0.768±0.028	0.854±0.007
	External data	0.587±0.088	0.741±0.017	0.555±0.112	0.746±0.017
XGBoost	Train data	0.882±0.003	0.861±0.002	0.746±0.008	0.877±0.002
	Test data	0.864±0.013	0.839±0.007	0.789±0.028	0.843±0.007
	External data	0.741±0.059	0.755±0.015	0.690±0.101	0.757±0.016

95% CI of the test data: 0.864±0.0131; AUC 95% CI of the external data: 0.7425±0.06) and XGBoost (AUC 95% CI of the test data: 0.8641±0.0129; AUC 95% CI of the external data: 0.7409±0.059) models performed relatively well. The ROC curve is shown in Figure 2A.

After fivefold cross-validation based on the complete data as a training set, integrated model prediction results were obtained and are shown in Table 5. The Bayes (AUC 95% CI of the test data: 0.8562±0.013; AUC 95% CI of the external data: 0.7989±0.0466) and logistic regression (AUC 95% CI of the test data: 0.8815±0.0118; AUC 95% CI of the external data: 0.7983±0.061) models were observed to not be more effective than the decision tree (AUC 95% CI of the test data: 0.8828±0.0117; AUC 95% CI of the external data: 0.814±0.0559), neural network (AUC 95% CI of the test data: 0.8826±0.0116; AUC 95% CI of the external data: 0.8083±0.0575), random forest (AUC 95% CI of the test data:

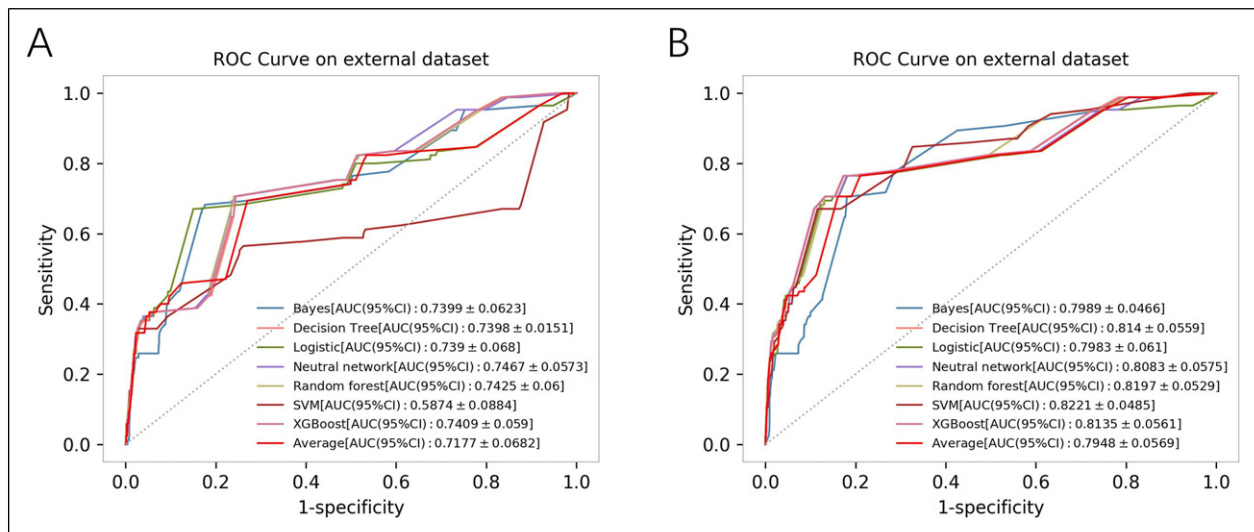
**Figure 2.** ROC curve analysis of different models based on the incomplete dataset (A) and complete dataset (B) for predicting metastasis in patients with RCC

Table 5. The Integrated Model Prediction Results After Five-Fold Cross-Validation in the Complete Data

Model	Set	Auc.	Acc.	Sens.	Spec.
Bayes	Train data	0.848±0.005	0.702±0.003	0.877±0.009	0.688±0.003
	Test data	0.856±0.013	0.830±0.006	0.728±0.031	0.838±0.007
	External data	0.799±0.047	0.817±0.014	0.644±0.101	0.822±0.014
Decision Tree	Train data	0.870±0.005	0.849±0.003	0.741±0.011	0.858±0.001
	Test data	0.883±0.012	0.840±0.006	0.791±0.026	0.844±0.007
	External data	0.814±0.056	0.842±0.013	0.701±0.104	0.847±0.014
Logistic	Train data	0.860±0.005	0.858±0.002	0.726±0.011	0.868±0.002
	Test data	0.882±0.012	0.837±0.007	0.792±0.026	0.840±0.007
	External data	0.798±0.061	0.834±0.014	0.701±0.104	0.837±0.014
Neutral network	Train data	0.870±0.005	0.848±0.003	0.743±0.011	0.856±0.003
	Test data	0.883±0.012	0.840±0.006	0.790±0.026	0.844±0.007
	External data	0.808±0.058	0.835±0.014	0.701±0.104	0.839±0.014
Random Forest	Train data	0.870±0.005	0.853±0.003	0.737±0.011	0.862±0.002
	Test data	0.882±0.012	0.838±0.006	0.795±0.026	0.841±0.007
	External data	0.820±0.053	0.841±0.013	0.701±0.104	0.846±0.014
SVM	Train data	0.833±0.006	0.848±0.002	0.737±0.011	0.857±0.003
	Test data	0.837±0.014	0.837±0.007	0.784±0.027	0.841±0.007
	External data	0.822±0.049	0.918±0.011	0.447±0.106	0.932±0.010
XGBoost	Train data	0.871±0.005	0.851±0.003	0.740±0.011	0.860±0.003
	Test data	0.882±0.012	0.840±0.006	0.789±0.026	0.844±0.007
	External data	0.814±0.056	0.841±0.013	0.701±0.104	0.846±0.014

0.8824±0.0118; AUC 95% CI of the external data: 0.8197±0.0529), SVM (AUC 95% CI of the test data: 0.8368±0.0142; AUC 95% CI of the external data: 0.8221±0.0485) and XGBoost (AUC 95% CI of the test data: 0.8823±0.0118; AUC 95% CI of the external data: 0.8135±0.0561) models. The ROC curve is shown in Figure 2B.

The model that was constructed with the complete training set outperformed the model that was constructed with the incomplete training set. Furthermore, the average prediction results of these two models based on the external test set were analyzed via the Delong test. The results revealed a p value of less than 0.0001, which demonstrated a significant difference in model performance.

Upsampling vs. Downsampling Analysis

Based on the analysis of positive and negative samples in each dataset, the disparity between the positive and negative samples significantly varied across the different datasets. Therefore, we investigated and compared the performances of the models that were constructed by using methods without sampling (the abovementioned results), along with methods involving upsampling and downsampling. We employed the fivefold cross-validation method to construct the models. First, downsampling was performed on the negative samples in the training set, and the model predictions after integrating the fivefold models were obtained, with the results being shown in [Supplementary Table 1](#) and Figure 3A. The SVM model achieved the highest accuracy, with an AUC 95% CI of 0.8685±0.0125 for the test data and 0.8274±0.0546 for the external data being reported. Afterward, upsampling was performed on the positive samples in the training set, and the model predictions after integrating the fivefold models were obtained, with the results being shown in [Supplementary Table 2](#) and Figure 3B. The XGBoost model demonstrated the highest accuracy, with an AUC 95% CI of 0.8819±0.0117 for the test data and 0.8162±0.0558 for the external data being reported. Both downsampling and upsampling led to relatively high AUC values and accuracy compared with those of the abovementioned nonsampling models.

To compare the performances of the three methods for constructing the models, a Delong test was also conducted on the average prediction results of these three models based on an external test set. The p value of the nonsampling model and the upsampling or downsampling model based on the external test set was less than 0.0001, thus indicating that the latter model performed better than the former model. The upsampling or downsampling method can significantly improve model performance. There was no significant difference observed in performance between the two models that were constructed by using the upsampling or downsampling methods (p=0.8734).

Based on the obtained results, we identified that the model constructed by using the upsampling method was the optimal model for this study. Furthermore, a visual analysis of the results is presented in [Figure 4](#) (DCA curve) and [Figure 5](#) (calibration curve). We also visualized each algorithm with a SHAP chart, as shown in [Supplementary Figure 1](#). SHAP

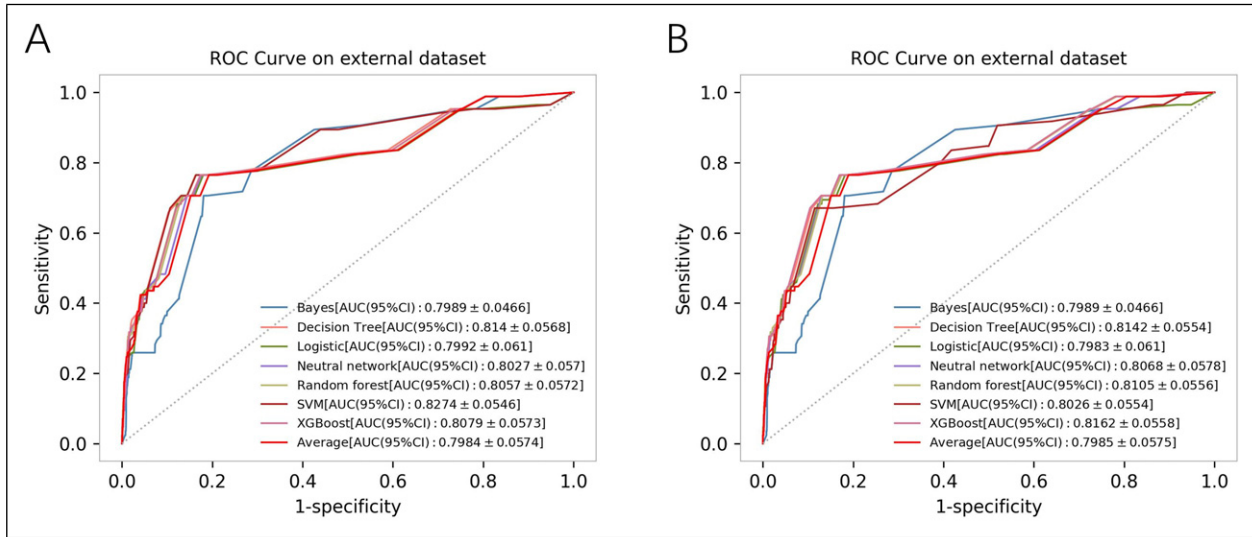


Figure 3. ROC curves based on the downsampling (A) and upsampling (B) analyses

summary plots for both the test and external datasets—generated using the full feature set (10 features)—were presented in [Supplementary Figure 2](#). Corresponding plots based on the reduced datasets (after removal of the top-ranked feature) were also provided for comparison ([Supplementary Figure 3](#)). Focusing on the full set of 10 features, we systematically evaluated the stability of feature rankings across top- n subsets (with n ranging from 3 to 8). As summarized in the [Supplementary Table 3](#), the results showed that while a subset of top-ranked features remains relatively consistent, the overall ranking order is notably sensitive to the removal of the highest-ranked feature. This instability reflects the label-

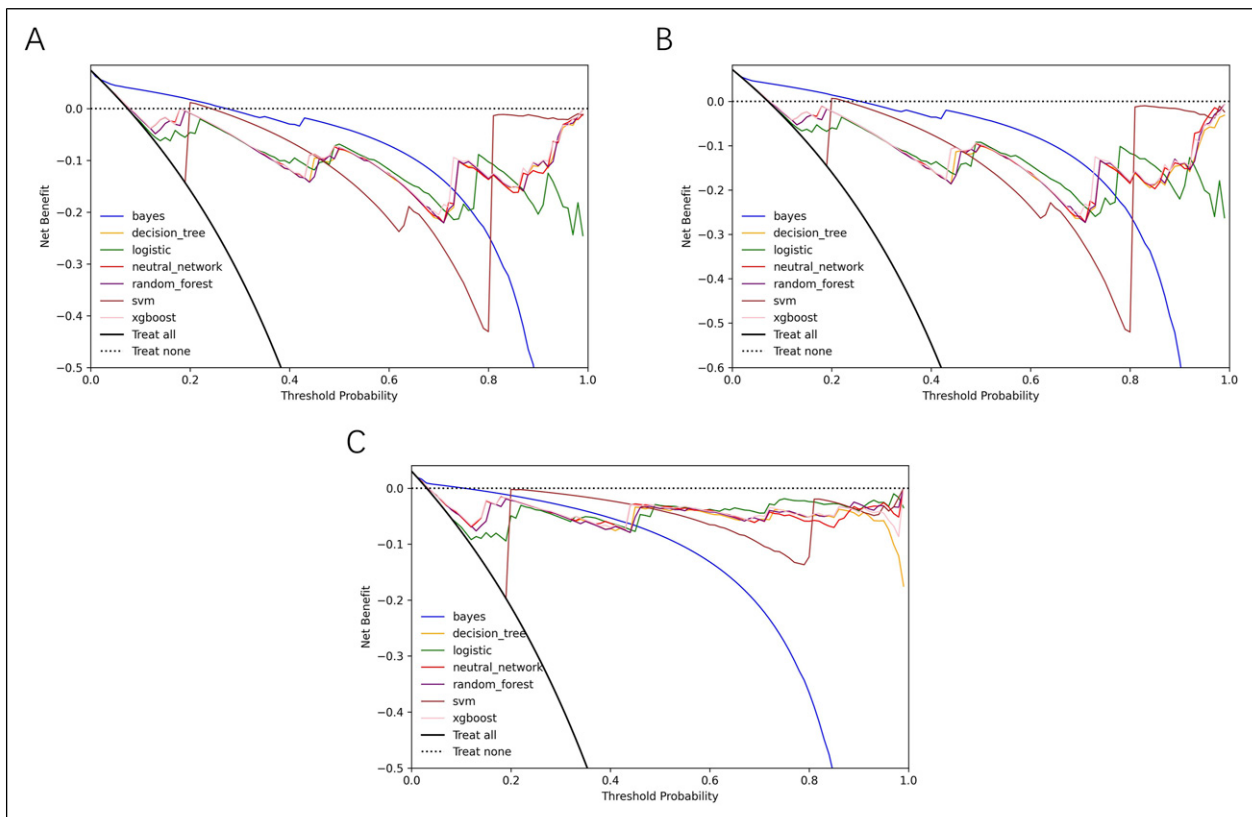


Figure 4. DCA curves based on the upsampling training set (A), test set (B), and external test set (C)

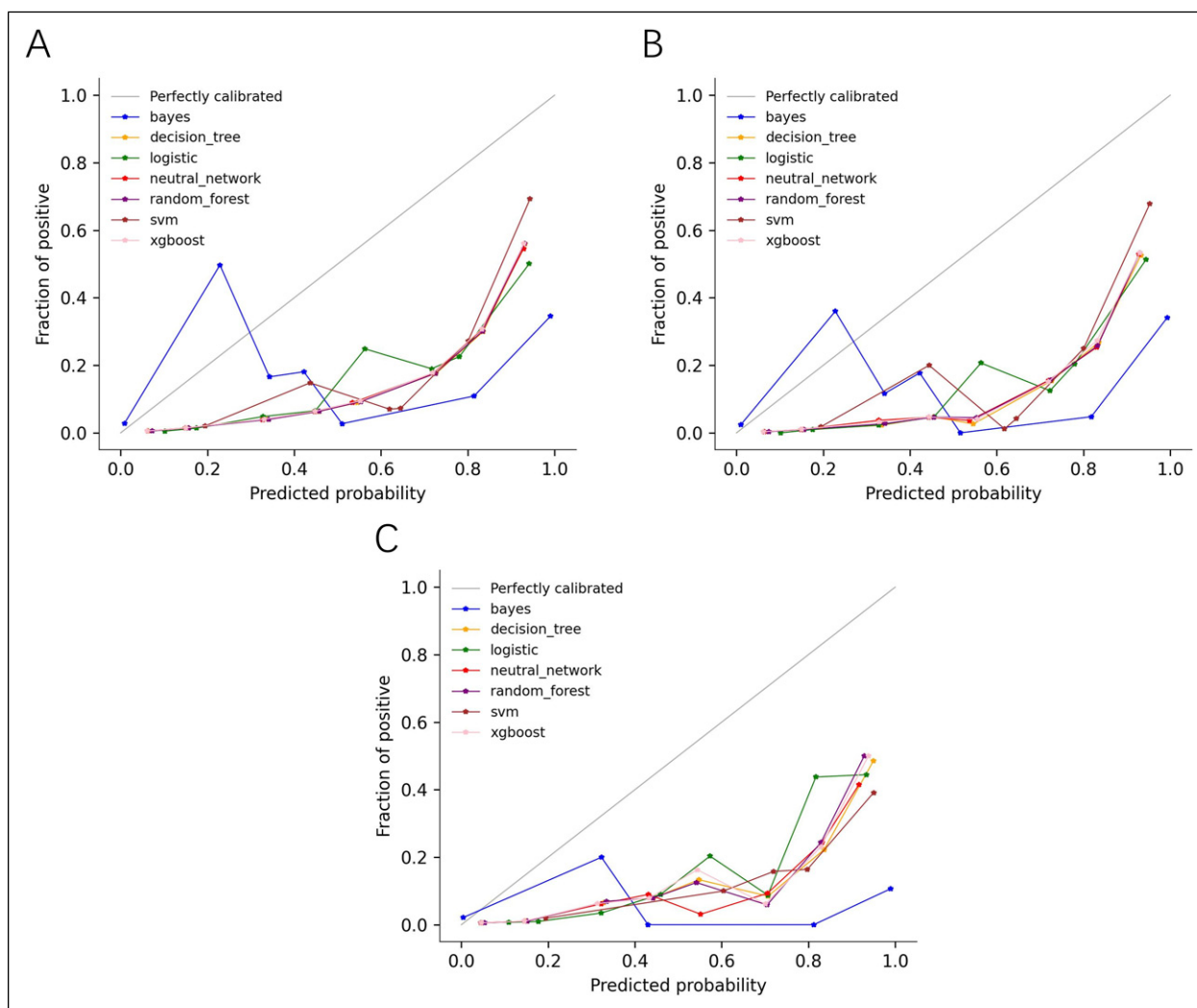


Figure 5. Calibration curves based on the upsampling training set (A), test set (B), and external test set (C)

driven optimization dynamics inherent in supervised learning models and highlights the need for caution when interpreting feature importance rankings as robust biological signals.

Based on the abovementioned feature visualization SHAP chart, it can be concluded that tumor size, T stage, N stage, and tumor grade demonstrate high feature importance in most models. Therefore, these four clinical characteristics can be considered as the main predictors of the metastasis parameters of RCC patients. Additionally, due to inherent imbalance of data, the model achieved low recall for the minority class, resulting in a substantially lower $F1(1)$ compared to $F1(0)$, as shown in [Supplementary Table 4](#).

Discussion

With respect to model development, although nomograms are currently the most commonly used prediction models, ML models are favored by an increasing number of medical workers because of their practicality, innovation, and accuracy.¹²⁻¹⁴ In clinical practice, tumor-related models can accurately predict prognosis by combining multiple factors, such as tumor pathological subtype, tumor stage, tumor diameter, and molecular marker expression. However, few researchers have attempted to use ML methods to explore the prediction of metastasis in RCC patients. This study mainly tried to apply ML method to model construction and prediction. Firstly, it is to test its ability of dealing with large amount of data. Secondly, it is to test its ability of integrating data. Thirdly, the weights of each feature of RCC can be roughly described. The results indicated good sensitivity and specificity. These algorithms can be applied to accurately predict whether RCC patients have metastasis, thereby providing assistance for the determination of clinical metastasis.

Various statistical measures can facilitate the understanding and interpretation of data. However, the limitations and efficiency of the processing of big data limit computing power and accuracy. In recent years, ML has included algorithmic methods that enable machines to solve problems without the use of specific computer programming, thereby providing an avenue for predictive modeling tasks.¹⁵⁻¹⁷ The integration of big data with ML algorithms is becoming a clinical necessity. In RCC studies, ML models can be applied to analyze the risk factors associated with specific diseases based on patient information. Yin et al¹⁸ integrated convolutional neural network models with Cox regression to identify potential prognostic biomarkers for overall survival. Terrematte P et al¹⁹ created a novel ML 13-gene signature, which was able to improve risk analysis and survival prediction for RCC patients. Chen S et al²⁰ developed and validated an ML-based prognosis prediction model, which could contribute to clinical decision-making for patients with RCC. Similarly, in our previous study, we investigated and demonstrated that ML algorithms could be used as auxiliary tools to predict the overall survival of patients with RCC using SEER data.¹¹

Once distant metastasis occurs in RCC, patient prognosis becomes very poor. Therefore, the prediction of distant metastasis in RCC patients is extremely important. Many clinical studies have used clinicopathological factors to establish models to predict metastatic risk in patients with RCC. Fan Z et al²¹ established nomograms using the SEER database to predict the risk of bone metastasis in patients with RCC. The calibration curve, ROC curve, and DCA confirmed good performance via diagnostic and prognostic nomograms. Wang J et al²² developed and validated a nomogram to predict distant metastasis in elderly patients with RCC. The AUC values of the training and validation cohorts indicated excellent predictive ability. DCA indicated that the clinical application value of the nomogram was better than that of traditional TN staging. Some scholars have also explored the use of ML algorithm models to make predictions. Xu C et al²³ used data from 40,355 RCC patients in the SEER database to construct an ML model for predicting the risk of bone metastasis in RCC patients. Among the prediction models established by the six ML algorithms, the XGBoost model achieved the best prediction performance (AUC = 0.891). Dong J and colleagues also used the SEER database to predict distant metastasis in RCC patients based on interpretable ML models.²⁴ The calibration curve indicated that the predicted values were highly consistent with the actual observed values.

The abovementioned studies indicate that ML models for predicting RCC metastasis are feasible and highly accurate. However, most of these studies were based on SEER data and lacked external validation. In contrast, in this study, we collected RCC data from the SEER database (training set) and the CNCC database (external test set). First, all of the data were preprocessed. The preprocessing of the feature data consisted of two steps. In step 1, clinicopathological data, including sex, tumor grade, side, pathological type, N stage, T stage, surgery and marital status, were subjected to numerical characterization through different number assignments. In step 2, for the two features of age and tumor size, the features in the training set and in the internal and external test sets were separately discretized. Second, we compared the performances of the models that were constructed via different ML algorithms for incomplete data and complete data. The accuracy of the ML models with complete datasets was observed to be significantly greater than that with incomplete datasets. Finally, to overcome the problem of large differences between positive and negative samples in the dataset, we also used upsampling and downsampling methods to reconstruct the model and test its performance. The results revealed that the accuracy of the model was further improved. Our study is the first to combine the SEER database with an external dataset to predict the distant metastasis of RCC. The results demonstrated that after validation with the external dataset, our ML model achieved high accuracy, thus providing considerable guiding value for clinical decision-making. If future applications require reliable individual-level probability estimates, post hoc calibration methods (such as Platt scaling or isotonic regression) should be incorporated. However, the “important features” identified here in this study should be interpreted strictly as contributors to predictive performance under the current modeling framework, not as evidence of biological causality.

This study has several limitations. First, the data obtained from the external validation cohort and the SEER database were retrospectively collected, which may introduce some inherent bias. In addition, to validate our prediction model in the general population, prospective clinical studies with larger sample sizes are necessary. The limited number of external patients may also have affected the statistical significance of the results. The underlying data distribution remains highly skewed (approximately 1:10~1:30 for positive: negative samples). Despite our best efforts, this extreme imbalance inherently limits the model's ability to achieve high recall for the minority class, resulting in a substantially lower F1(1) compared to F1(0). This work represents an initial exploratory attempt, and future studies should incorporate more advanced imbalance-handling techniques or larger cohorts. Second, we were unable to obtain biomarkers, blood test results or time to the development of metastasis from the SEER database. In addition, feature importance reflects the predictive contribution within the model and should not be directly equated with the true clinical weight of risk factors. We expect that the addition of data from external data validation will result in a more sophisticated and effective adoption of ML models as supplementary tools for prediction research.

Conclusion

In summary, using data from the SEER database and CNCC dataset, this study explored the factors related to the prediction of RCC metastasis through multi-algorithm ML models. Relevant algorithms were established to predict the possibility of distant metastasis in RCC patients. The findings demonstrate that ML algorithms can effectively predict distant metastasis in patients with RCC and play a positive role in clinical applications.

Acknowledgements

We are especially grateful to Moxuan Wang, who is a 14-year-old girl. She helped us with the layout and beautification of the figures. To encourage, she was listed one of authors.

ORCID iD

Li Wen  <https://orcid.org/0009-0008-1425-3435>

Ethical Considerations

The authors state that they have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. Our study was approved by Institutional Review committee of the National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences (NCC/CHCAMS) (Institutional Review Board number: 21/405-3076, date: October 13, 2021).

Consent for Publication

Patient study consent was not required due to the study's retrospective nature. The requirement for obtaining informed consent was waived by the Institutional Review committee.

Author Contributions

All authors listed in this manuscript contributed significantly to the study. Yajian Li and Xinwei Wang contributed to writing the manuscript. Moxuan Wang contributed to layout and beautification of the figures. Jianzhong Shou contributed to supervision. Li Wen and Cancan Chen contributed to reviewing the manuscript for critical revisions. All authors read and approved the final manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The datasets used and/or analyzed data in the current study are available from the corresponding author on reasonable request.

Supplemental Material

Supplemental material for this article is available online.

References

1. National Cancer Center; Renal Cancer Expert Committee of National Cancer Quality Control Center. Quality control index for standardized diagnosis and treatment of renal cancer in China (2022 edition). *Zhonghua Zhong Liu Za Zhi*. 2022;44:1256-1261.
2. Motzer RJ, Jonasch E, Agarwal N, et al. NCCN Guidelines® Insights: Kidney Cancer, Version 2.2024. *J Natl Compr Canc Netw*. 2024;22:4-16.
3. Tannir NM, Albigès L, McDermott DF, et al. Nivolumab plus ipilimumab versus sunitinib for first-line treatment of advanced renal cell carcinoma: extended 8-year follow-up results of efficacy and safety from the phase III CheckMate 214 trial. *Ann Oncol*. 2024; 35:1026-1038.
4. Motzer RJ, Porta C, Eto M, et al. Lenvatinib Plus Pembrolizumab Versus Sunitinib in First-Line Treatment of Advanced Renal Cell Carcinoma: Final Prespecified Overall Survival Analysis of CLEAR, a Phase III Study. *J Clin Oncol*. 2024;42:1222-1228.
5. Powles T, Albigès L, Bex A, et al. Renal cell carcinoma: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol*. 2024;35:692-706.

6. Bex A, Ghanem YA, Albiges L, et al. European Association of Urology Guidelines on Renal Cell Carcinoma: The 2025 Update. *Eur Urol.* 2025;87:683-696.
7. Rathmell WK, Rumble RB, Van Veldhuizen PJ, et al. Management of Metastatic Clear Cell Renal Cell Carcinoma: ASCO Guideline. *J Clin Oncol.* 2022;40:2957-2995.
8. Rose TL, Kim WY. Renal Cell Carcinoma: A Review. *JAMA.* 2024;332:1001-1010.
9. Lee CH, Kang M, Kwak C, et al. Sites of Metastasis and Survival in Metastatic Renal Cell Carcinoma: Results From the Korean Renal Cancer Study Group Database. *J Korean Med Sci.* 2024;39:e293.
10. Dogan I, Iribas A, Paksoy N, Vatansever S, Basaran M. Outcomes and prognostic factors in metastatic renal cell carcinoma patients with brain metastases. *J Cancer Res Ther.* 2023;19:S587-S591.
11. Jiang W, Chen Z, Chen C, Wang L, Han T, Wen L. Machine learning algorithms being an auxiliary tool to predict the overall survival of patients with renal cell carcinoma using the SEER database. *Transl Androl Urol.* 2024;13:53-63.
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378.
13. Wang X, Zhao J, Marostica E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature.* 2024;634:970-978.
14. Feng Y, Long Y, Wang H, et al. Benchmarking machine learning methods for synthetic lethality prediction in cancer. *Nat Commun.* 2024;15:9058.
15. MacEachern SJ, Forkert ND. Machine learning for precision medicine. *Genome.* 2021;64:416-425.
16. Le NQ, Ou YY. Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins. *BMC Bioinformatics.* 2016;17(Suppl 19):501.
17. Le NQ, Nguyen TT, Ou YY. Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties. *J Mol Graph Model.* 2017;73:166-178.
18. Yin Q, Chen W, Zhang C, Wei Z. A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection. *Lab Investig.* 2022;102:1064-1074.
19. Terrematte P, Andrade DS, Justino J, Stransky B, de Araújo DSA, Dória Neto AD. A Novel Machine Learning 13-Gene Signature: Improving Risk Analysis and Survival Prediction for Clear Cell Renal Cell Carcinoma Patients. *Cancers (Basel).* 2022;14:2111.
20. Chen S, Guo T, Zhang E, et al. Machine learning-based prognosis signature for survival prediction of patients with clear cell renal cell carcinoma. *Heliyon.* 2022;8:e10578.
21. Fan Z, Huang Z, Huang X. Bone Metastasis in Renal Cell Carcinoma Patients: Risk and Prognostic Factors and Nomograms. *J Oncol.* 2021;2021:5575295.
22. Wang J, Zhanghuang C, Tan X, et al. Development and Validation of a Nomogram to Predict Distant Metastasis in Elderly Patients With Renal Cell Carcinoma. *Front Public Health.* 2022;9:831940.
23. Xu C, Liu W, Yin C, et al. Establishment and Validation of a Machine Learning Prediction Model Based on Big Data for Predicting the Risk of Bone Metastasis in Renal Cell Carcinoma Patients. *Comput Math Methods Med.* 2022;2022:5676570.
24. Dong J, Duan M, Liu X, et al. Prediction of Distant Metastasis of Renal Cell Carcinoma Based on Interpretable Machine Learning: A Multicenter Retrospective Study. *J Multidiscip Healthc.* 2025;18:195-207.