



## ORIGINAL ARTICLE

# Machine learning-based prognostic and metastasis models of kidney cancer

Yuxiang Zhang<sup>1</sup>  | Na Hong<sup>2</sup>  | Sida Huang<sup>3</sup> | Jie Wu<sup>1</sup> | Jianwei Gao<sup>2</sup> | Zheng Xu<sup>2</sup> | Fubo Zhang<sup>1</sup> | Shaohui Ma<sup>1</sup> | Ye Liu<sup>1,4</sup> | Peiyuan Sun<sup>1</sup> | Yanping Tang<sup>1</sup> | Chun Liu<sup>2</sup> | Jianzhong Shou<sup>1</sup> | Meng Chen<sup>1</sup>

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

<sup>2</sup>Digital Health China Technologies, Co., Ltd., Beijing, China

<sup>3</sup>Department of public policy, Cornell University, Ithaca, New York, USA

<sup>4</sup>The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital/National Center of Gerontology of National Health Commission, Beijing, China

## Correspondence

Jianzhong Shou and Meng Chen, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China.  
Email: [shoujianzhong@cicams.ac.cn](mailto:shoujianzhong@cicams.ac.cn) and [chenmeng@cicams.ac.cn](mailto:chenmeng@cicams.ac.cn)

## Funding information

CAMS Innovation Fund for Medical Sciences (CIFMS), Grant/Award Number: 2021-I2M-1-066; Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences,

## Abstract

**Background:** Kidney cancer originates from the urinary tubule epithelial system of the renal parenchyma, accounting for 20% of all urinary system tumors. Approximately 70% of cases are localized at diagnosis, and 30% are metastatic. Most localized kidney cancers can be cured by surgery, but most metastatic patients relapse after surgery and eventually die of kidney cancer. Therefore, accurately predicting patient survival and identifying high-risk metastatic patients will effectively guide interventions and improve prognosis.

**Methods:** This study used the data of 12,394 kidney cancer patients from the surveillance, epidemiology, and end results database to construct a research cohort related to kidney cancer survival and metastasis. Eight machine learning models (including support vector machines, logistic regression, decision tree, random forest, XGBoost, AdaBoost, K-nearest neighbors, and multilayer perceptron) were developed to predict the survival and metastasis of kidney cancer and six evaluation indicators (accuracy, precision, sensitivity, specificity, F1 score, and area under the receiver operating characteristic [AUROC]) were used to verify, evaluate, and optimize the models.

**Results:** Among the eight machine learning models, Logistic Regression has the highest AUROC in both prediction scenarios. For 3-year survival prediction, the Logistic Regression model had an accuracy of 0.684, a sensitivity of 0.702, a specificity of 0.670, a precision of 0.686, an F1 score of 0.683, and an AUROC of 0.741. For tumor metastasis prediction, the Logistic Regression model had an

**Abbreviations:** AJCC, American joint committee on cancer; AUROC, the area under the receiver operating characteristic; DCCPS, NCI's Division of Cancer Control and Population Sciences; IMDC, International Metastatic Renal-Cell Carcinoma Database Consortium; KNN, K-nearest neighbors; LOH, loss of heterozygosity; MSKCC, Memorial Sloan-Kettering Cancer Center; ROC, receiver operating characteristic; SEER, Surveillance, Epidemiology, and End Results; SRP, surveillance research program; SVM, support vector machines; UISS, Los Angeles integrated staging system.

[Correction added on 13 August 2022, after first publication: The missing funding information was included.]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Cancer Innovation* published by John Wiley & Sons Ltd. on behalf of Tsinghua University Press.

Grant/Award Number: 2019PT320027;  
Beijing Hope Run Special Fund of Cancer  
Foundation of China,  
Grant/Award Number: LC2019A04;  
Fundamental Research Funds for the  
Central Universities,  
Grant/Award Number: 3332020023

accuracy of 0.800, a sensitivity of 0.540, a specificity of 0.830, a precision of 0.769, an F1 score of 0.772, and an AUROC of 0.804.

**Conclusion:** In this study, we selected appropriate variables from both statistical and clinical significance and developed and compared eight machine learning models for predicting 3-year survival and metastasis of kidney cancer. The prediction results and evaluation results demonstrated that our model could provide decision support for early intervention for kidney cancer patients.

#### KEYWORDS

machine learning, kidney cancer, survival, metastasis, prognostic model

## 1 | BACKGROUND

Kidney cancer is a malignant tumor that originates from the urinary tubule epithelial system of the renal parenchyma, accounting for approximately 20% of all urinary system tumors [1]. As of 2018, the number of new cases of kidney cancer worldwide exceeded 400,000, and the number of deaths exceeded 170,000 [2]. In China, the annual incidence of kidney cancer is increasing. In 2016, the incidence of kidney cancer exceeded 15,000, and deaths exceeded 5000. The incidence rate was 4.02 per 100,000, and the mortality rate was 1.37 per 100,000 [3]. The surveillance, epidemiology, and end results (SEERs) Program provides information on cancer statistics to reduce the cancer burden among the US population. This program was supported by the surveillance research program (SRP) of the NCI's Division of Cancer Control and Population Sciences (DCCPS) [4]. The data from the SEER database is collected from the cancer registry of the population and currently covers 30% of the population data in the United States. Globally, about half of kidney cancer cases are diagnosed at age 65, with peak incidence at age 75 [5]. Early-stage localized kidney cancer have a good prognosis through radical nephrectomy, but patients with stage II or III have a high risk of recurrence after nephrectomy [5]. Immunotherapy has become the most promising treatment for patients with metastatic kidney cancer, as kidney cancer is resistant to chemotherapy [6, 7]. But immunotherapy is only successful in 10%–15% of patients [7]. Therefore, treatment for metastatic kidney cancer remains inadequate.

Renal cell carcinoma is the most common malignant renal tumor. Risk prediction models such as Memorial Sloan-Kettering Cancer Center (MSKCC), International Metastatic Renal-Cell Carcinoma Database Consortium (IMDC), Leibovich, University of California, and Los Angeles Integrated Staging System (UISS) have been applied to identify patients at high risk of recurrence after surgery [8]. Traditional statistical analysis is a basic description of sample

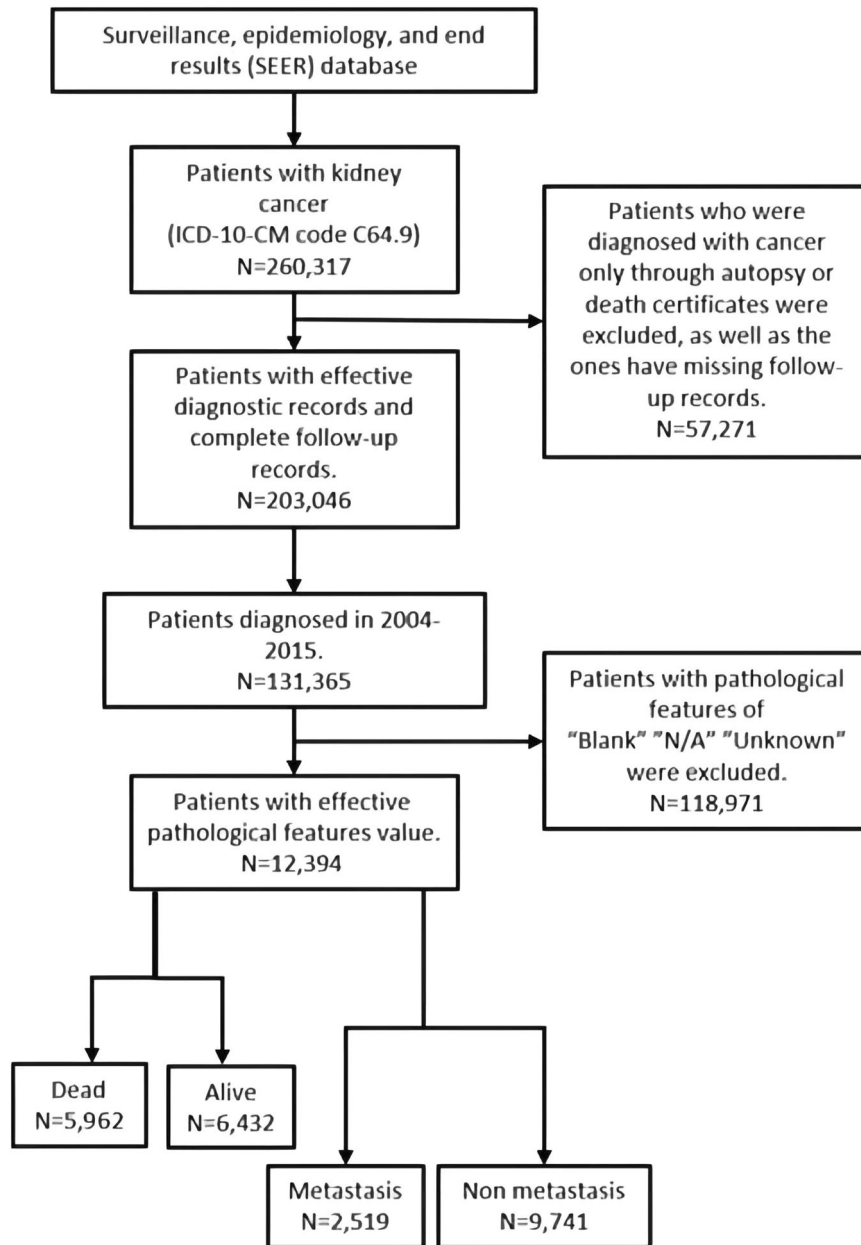
data based on assumption of sample distribution, and then make inferences in the form of probability, and it is mainly inferred through the observation of a random process in a limited period of time. Although the predictive capabilities have been improved, it is still far from meeting the clinical needs for patient classification accurately. In contrast, based on rich data, machine learning methods predict future unknown outcomes by generalizing features without making assumptions about data distribution or understanding the mechanisms behind sample data. Due to the complexity and diversity of a large amount of cancer data, traditional statistical inference models are facing challenges when processing and analyzing it effectively or efficiently; however, machine learning methods demonstrated advantages for analyzing huge volume and high complexity data. For example, Byun et al. used deep learning method to predict the prognosis in nonmetastatic clear cell renal cell carcinoma [9]. Machine learning models also showed good performance in some survival prediction analyses for other cancer types. Gu-Wei Ji et al. used machine learning models to predict the recurrence of hepatocellular carcinoma after resection [10].

Therefore, we incorporate machine learning models into this study to explore a more efficient, intelligent, and accurate prediction model for the prognosis and metastasis risk assessment of kidney cancer patients.

## 2 | MATERIALS AND METHODS

### 2.1 | Patient data set

We included patients with histologically diagnosed kidney cancer who had complete survival time and active follow-up data from 2004 to 2015 in the SEER database. The data selection process was illustrated in Figure 1. Kidney and renal pelvic cancers were described together in the SEER database. We chose the disease with ICD-10-CM code C64.9: Malignant neoplasm of the unspecified kidney, except the



**FIGURE 1** The process of data selection

renal pelvis. Patients diagnosed with cancer only through autopsy or death certificates, patients had missing follow-up records and patients with pathological features of “Blank” “N/A” “Unknown” were excluded. Tumor staging was coded according to the sixth edition of the TNM staging system of the American Joint Committee on Cancer of the United States.

## 2.2 | Data preprocessing and feature selection

The observation period for all patients was from the date of diagnosis to death, recurrence, or the end of data inclusion

period (2018). Age at the first visit, race (black, white, other, unknown), sex, tumor size, marital status at diagnosis, year of birth, year of diagnosis, histologic type, Fuhrman nuclear grade (I, II, III, and IV), tumor stage (T1, T2, T3, T4, and Unknown), lymph node status (N0, N1, N2, and Unknown), distant metastasis (M0, M1, and Unknown), primary site surgery information, bone metastases, brain metastases, liver metastases, lung metastases, overall survival time, and survival status were recorded. The American Joint Committee on Cancer Staging Manual (6th edition) was used in this study. The primary endpoints of the study were whether the patient died within 36 months of diagnosis and whether tumor had metastasis within 3 years after diagnosis. Data preprocessing, including data

cleaning, data integration, and data transformation, was performed. All variables included in analysis were reviewed by clinicians, we excluded variables that are irrelevant with the outcomes. The  $\chi^2$  test was used to assess the difference between categorical variables, and the *t*-test was used to assess the difference between continuous variables. Differences were considered statistically significant at  $p < 0.05$ . All analyses were performed using R (version 4.2.0) and R Studio (version 1.3.1093) software.

### 2.3 | Development and evaluation of prognostic models

We used 80% of the data as training set for model development and the rest 20% of the data as testing set. Eight machine learning models were applied to predict the outcomes including support vector machines (SVMs, also support-vector networks) [11], logistic regression [12], decision trees [13], random forests [14], XGBoost [15], Adaptive Boosting (AdaBoost) [16], K-nearest neighbors (KNN) [17], and multilayer perceptions (MLP) [18].

Machine learning models were developed using Scikit-learn (version 0.23.2) and we applied 5-fold cross-validation to improve the stability of the model. We compared and evaluated the model performance using six indicators as follows: Accuracy is a measure of how close the observed value (measured value) is to the true value of the quantity. If error is less, the measurement is accurate [19]. Precision is the fraction of relevant instances among the retrieved instances [20]. Recall (sensitivity) is the fraction of relevant instances that were retrieved [20]. Specificity refers to the proportion of those who do not have the condition (when judged by the “Gold Standard”) that received a negative result on this test [21]. F1 score is the harmonic mean of the precision and recall [20]. Area under the receiver operating characteristic curve (AUC) analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and before specifying) the cost context or the class distribution. Receiver operating characteristic (ROC) analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision-making [22].

## 3 | RESULTS

### 3.1 | Clinical characteristics of patients

Data of 12,394 eligible patients with kidney cancer from the SEER database from 2004 to 2015 were extracted. Among them, 6432 (51.90%) patients survived for more

than 3 years, and 2519 (20.32%) patients had metastases. The distributions between clinicopathological features and 3-year overall survival rate were summarized in Table 1. And the distributions between clinicopathologic features and metastasis were listed in Table 2. The median age of the participants was 61 (49–73) years, accounting for 82.35% of the patients, and 65.84% of the patients were male. The most common histological subtype was clear cell renal cell carcinoma (59.20%). The number of patients in the localized and regional stages was 41.01% and 37.97%, respectively. The follow-up period for this cohort was until 2018.

### 3.2 | Data preprocessing and feature selection

After data preprocessing, a total of 12,082 kidney cancer patients were selected for the analysis of survival, and 12,192 cases were selected for the analysis of metastasis.

Considering the statistical significance and the clinical relevance from clinicians' review, we finalized the variables for subsequent machine learning modeling. The variables for predicting 3-year survival were race, age of diagnosis, tumor size, differentiation grade, stage, histologic type, TNM staging, primary tumor surgery, and lymph node clearance. The variables for predicting metastasis were race, gender, age of diagnosis, grade, histologic type, and T and N stage.

### 3.3 | Performance of prognostic models

Eight machine learning models including SVM, logistic regression, decision tree, random forest, XGBoost, AdaBoost, KNN, and multilayer perceptron were developed to predict the survival and metastasis of kidney cancer patients. Accuracy, precision, recall (sensitivity), specificity, F1 score, and area under the ROC curve (AUC) were applied used to evaluate the model performance. The results showed that logistic regression had the best performance in terms of AUC of 0.741 for 3-year survival and AUC of 0.804 for metastasis, respectively (Figures 2 and 3). For the 3-year survival predictive study, logistic regression model had the accuracy of 0.684, the sensitivity of 0.702, the specificity of 0.670, the precision of 0.686, and the F1 score of 0.683. For the tumor metastasis predictive study, the Logistic regression model had the accuracy of 0.800, the sensitivity of 0.540, the specificity of 0.830, the precision of 0.769, and the F1 score of 0.772 (Tables 3 and 4). When we stratified the population by histological type, the patients with clear cell renal carcinoma demonstrated the similar findings

**TABLE 1** Relationship between clinicopathologic features and 3-year overall survival rate

Characteristics	Total <i>n</i> = 12,394	3-year alive <i>n</i> = 6432	3-year death <i>n</i> = 5962	<i>p</i> -value
<b>Gender <i>n</i> (%)</b>				0.302
Male	8160 (65.84%)	4207 (65.41%)	3953 (66.3%)	
Female	4234 (34.16%)	2225 (34.59%)	2009 (33.7%)	
<b>Age(mean ± SD)</b>	60.67 ± 12.29	59.77 ± 12.13	61.65 ± 12.39	<i>p</i> < 0.001
<b>Race <i>n</i> (%)</b>				0.016
White	10,207 (82.35%)	5305 (82.48%)	4902 (82.22%)	
Black	1304 (10.52%)	669 (10.40%)	635 (10.65%)	
Other	814 (6.57%)	423 (6.58%)	391 (6.56%)	
Unknown	69 (0.56%)	35 (0.54%)	34 (0.57%)	
<b>Tumor size (mm, mean ± SD)</b>	92.81 ± 105.99	83.61 ± 93.56	102.67 ± 117.16	<i>p</i> < 0.001
<b>Grade <i>n</i> (%)</b>				<i>p</i> < 0.001
Well differentiated; Grade I	674 (5.44%)	506 (7.87%)	168 (2.82%)	
Moderately differentiated; Grade II	4142 (33.42%)	2727 (42.40%)	1415 (23.73%)	
Poorly differentiated; Grade III	4872 (39.31%)	2467 (38.36%)	2405 (40.34%)	
Undifferentiated; anaplastic; Grade IV	2706 (21.83%)	732 (11.38%)	1974 (33.11%)	
<b>Histologic. Type. ICD-O-3 (n(%))</b>				<i>p</i> < 0.001
Clear cell renal cell carcinoma	7337 (59.20%)	4027 (62.62%)	3310 (55.52%)	
Others	5057 (40.80%)	2405 (37.39%)	2652 (44.48%)	
<b>Stage <i>n</i> (%)</b>				<i>p</i> < 0.001
Localized	5083 (41.01%)	3630 (56.44%)	1453 (24.37%)	
Regional	4706 (37.97%)	2257 (35.09%)	2449 (41.08%)	
Distant	2605 (21.02%)	545 (8.47%)	2060 (34.55%)	
<b>AJCC T 6th <i>n</i> (%)</b>				<i>p</i> < 0.001
T0	2 (0.02%)	0 (0.00%)	2 (0.03%)	
T1 (T1a, T1b, T1NOS)	3663 (29.56%)	2500 (38.87%)	1163 (19.50%)	
T2	2217 (17.89%)	1422 (22.11%)	795 (13.33%)	
T3 (T3a, T3b, T3NOS)	5838 (47.11%)	2378 (36.97%)	3460 (58.02%)	
T4	617 (4.98%)	112 (1.74%)	505 (8.47%)	
Unknown	57 (0.46%)	20 (0.31%)	37 (0.62%)	
<b>AJCC N 6th <i>n</i> (%)</b>				<i>p</i> < 0.001
N0	9532 (76.91%)	5757 (89.51%)	3775 (63.31%)	
N1	1641 (13.24%)	407 (6.33%)	1234 (20.70%)	
N2	1127 (9.09%)	211 (3.28%)	916 (15.36%)	
Unknown	94 (0.76%)	57 (0.89%)	37 (0.62%)	
<b>AJCC M 6th <i>n</i> (%)</b>				<i>p</i> < 0.001
M0	9741 (78.59%)	5833 (90.69%)	3908 (65.55%)	
M1	2519 (20.33%)	518 (8.05%)	2001 (33.56%)	
Unknown	134 (1.08%)	81 (1.26%)	53 (0.89%)	

TABLE 1 (Continued)

Characteristics	Total n = 12,394	3-year alive n = 6432	3-year death n = 5962	p-value
<b>Primary site surgery information n (%)</b>				<i>p</i> < 0.001
None	80 (0.65%)	6 (0.09%)	74 (1.24%)	
Partial nephrectomy	1035 (8.35%)	650 (10.11%)	385 (6.46%)	
Radical nephrectomy	11,138 (89.87%)	5697 (88.57%)	5441 (91.26%)	
Surgery	131 (1.05%)	72 (1.12%)	59 (0.99%)	
Unknown	10 (0.08%)	7 (0.11%)	3 (0.05%)	
<b>Regional nodes examined n (%)</b>				<i>p</i> < 0.001
None	381 (3.07%)	189 (2.94%)	192 (3.22%)	
lymph node dissection	11,894 (95.97%)	6164 (95.83%)	5730 (96.11%)	
No record	119 (0.96%)	79 (1.23%)	40 (0.67%)	

Abbreviation: AJCC, American joint committee on cancer.

(Table 5). For the 3-year survival, logistic regression model had the AUC of 0.710, the accuracy of 0.658, the sensitivity of 0.674, the specificity of 0.649, the precision of 0.661, and the F1 score of 0.653. For the metastasis, logistic regression model had the AUC of 0.811, the accuracy of 0.826, the sensitivity of 0.593, the specificity of 0.851, the precision of 0.802, and the F1 score of 0.803.

## 4 | DISCUSSION

The incidence of kidney cancer has been increasing annually. Renal cell carcinoma is the most common type of kidney cancer. Most localized renal cell carcinomas had a good prognosis after cytoreductive nephrectomy, but approximately 20%–30% of patients died due to the recurrence and metastasis after surgery. Precisely prediction the clinical outcome for early diagnosis and treatment are essential to improve the prognosis of kidney cancer.

Clear cell renal cell carcinoma (ccRCC) accounts for 60%–80% of all renal cell carcinomas [23]. ccRCC has a better clinical prognosis than other types of kidney cancer. In our multivariate logistic regression model for survival and metastasis, the ccRCC patients showed better outcomes with an odds ratio 0.897 (95% confidence interval: 0.825–0.975, *p* = 0.01) for survival and an odds ratio 1.372 (95% confidence interval: 1.235–1.525, *p* < 0.001) for metastasis compared to other histological types. ccRCC is characterized by loss of chromosome 3p and biallelic inactivation of the *VHL* gene due to mechanisms such as loss of heterozygosity (LOH), mutation, and methylation [24]. High-frequency mutations of genes (*PBRM1*, *SETD2*, *BAP1*, etc.) involved in epigenetic regulation were also found in more than 50%

of ccRCC [24]. In this –omics era, biomarkers especially the genetic biomarkers has been widely applied in clinical area. The genetic signature has been incorporated in risk prediction models, such as ClearCode34 and 16-gene assay [25, 26]. In the future, the genetic biomarkers should be incorporated in our current model and be evaluated in an independent population.

Clinical prediction models are commonly used to help clinicians and patients make better medical decisions, and to help government departments and health managers better manage medical quality and rationally allocate medical resources. In recent years, with the advance of science and technology, cancer diagnosis and prediction has entered a multidimension era of big data. The conventional clinical data measurements include images, genes, blood, proteins, pathological analysis, etc., which bring great challenges to the data analysis and disease diagnosis, and due to various reasons, some data will be partially missing. At present, the methods of cancer prediction are mainly divided into traditional statistical inference and machine learning models. The former is the basic description of the sample data based on the assumption of the sample distribution, and then inferences in the form of probabilities. And it is mainly inferred by observing a random process in a limited period of time. Machine learning models are very different from statistical inferences in that they take advantage of rich data and learn new patterns through generalized features to predict future outcomes without the need to make assumptions about the distribution of the data and understand the mechanisms behind the sample data. Due to the complexity and diversity of large amounts of data, traditional statistical inference models are less accurate than machine learning models,

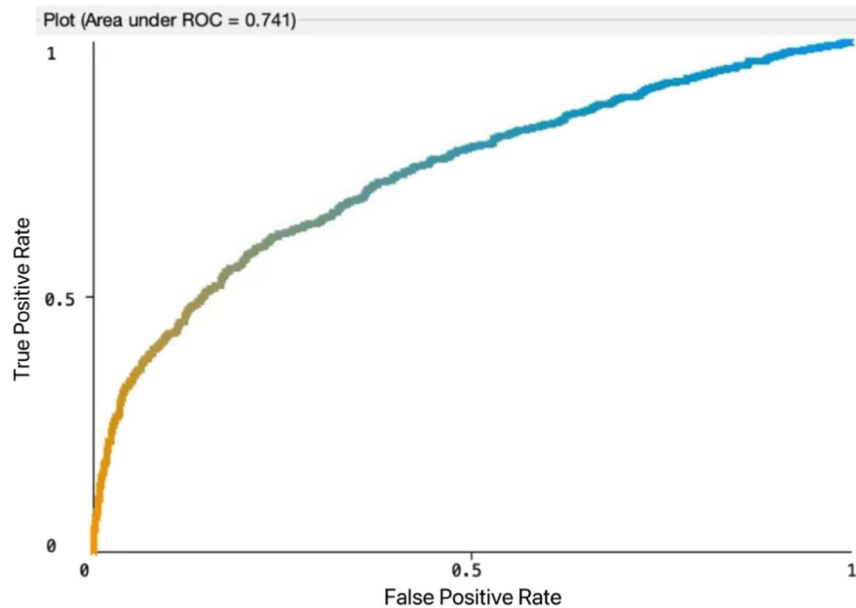
**TABLE 2** Relationship between clinicopathologic features and metastasis

Characteristics	Metastasis <i>n</i> = 2519	No metastasis <i>n</i> = 9741	<i>p</i> -value
<b>Gender <i>n</i> (%)</b>			0.003
Male	1722 (68.36%)	6350 (65.19%)	
Female	797 (31.64%)	3391 (34.81%)	
<b>Age (mean ± SD)</b>	59.69 ± 11.48	60.93 ± 12.48	<i>p</i> < 0.001
<b>Race <i>n</i> (%)</b>			0.007
White	2115 (83.96%)	7984 (81.96%)	
Black	221 (8.77%)	1078 (11.07%)	
Other	176 (6.99%)	618 (6.34%)	
Unknown	7 (0.28%)	61 (0.63%)	
<b>Tumor size (mm, mean ± SD)</b>	119.10 ± 130.77	85.81 ± 95.89	<i>p</i> < 0.001
<b>Grade <i>n</i> (%)</b>			<i>p</i> < 0.001
Well differentiated; Grade I	29 (1.15%)	624 (6.41%)	
Moderately differentiated; Grade II	325 (12.90%)	3774 (38.74%)	
Poorly differentiated; Grade III	1063 (42.20%)	3765 (38.65%)	
Undifferentiated; anaplastic; Grade IV	1102 (43.75%)	1578 (16.20%)	
<b>Histologic type ICD-O-3 <i>n</i> (%)</b>			0.001
Clear cell renal cell carcinoma	1419 (60.07%)	5851 (56.33%)	
Others	1100 (39.93%)	3890 (43.67%)	
<b>AJCC T 6th <i>n</i> (%)</b>			<i>p</i> < 0.001
T0	1 (0.04%)	1 (0.01%)	
T1 (T1a, T1b, T1NOS)	161 (6.39%)	3458 (35.50%)	
T2	284 (11.27%)	1922 (19.73%)	
T3 (T3a, T3b, T3c, T3NOS)	1709 (67.85%)	4066 (41.74%)	
T4	333 (13.22%)	270 (2.77%)	
Unknown	31 (1.23%)	24 (0.25%)	
<b>AJCC N 6th <i>n</i> (%)</b>			<i>p</i> < 0.001
N0	1200 (47.64%)	8272 (84.92%)	
N1	688 (27.31%)	930 (9.55%)	
N2	606 (24.06%)	506 (5.19%)	
Unknown	25 (0.99%)	33 (0.34%)	

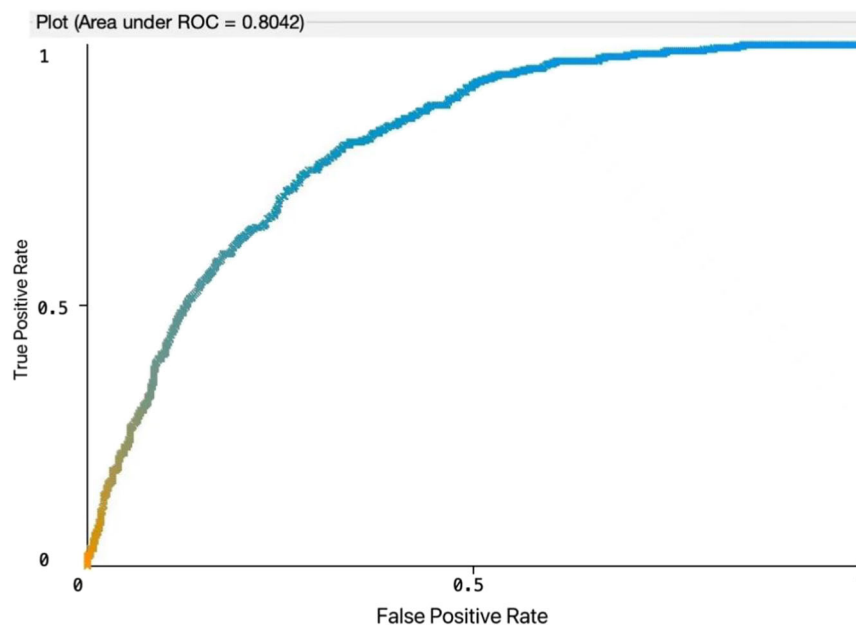
Abbreviation: AJCC, American joint committee on cancer.

and the general applicability is not as strong as the latter, so modern cancer predictions are gradually incorporated into machine learning models. Machine learning also has its inherent problems, such as the inability to judge the advantages and disadvantages of the model from the algorithm, and the results are difficult to interpret, so these problems need to be evaluated and overcome through external verification before actual use.

Eight machine learning models were applied in this study. SVMs, also support-vector networks are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis [11]. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of



**FIGURE 2** ROC curve of the optimal model logistic regression on 3-year survival prediction



**FIGURE 3** ROC curve of the optimal model logistic regression on tumor metastasis prediction

binary regression) [12]. Decision trees and closely related influence diagrams are used as visualization and analysis decision support tools, in which the expected value (or expected utility) of competing solutions is calculated [13]. Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time [14]. XGBoost aims to provide a “scalable, portable, and distributed gradient boosting (GBM, GBRT, and GBDT) library.” It runs on a single machine, along with distributed processing frameworks

Apache Hadoop, Apache Spark, Apache Flink, and Dash [15]. AdaBoost, is a statistical classification meta-algorithm that can be used in conjunction with many other types of learning algorithms to improve performance, and often referred to as the best out-of-the-box classifier [16]. K-nearest neighbors (KNN) is a nonparametric classification method used for classification and regression. In both cases, the input consists of the  $k$  training examples closest to the data set. The output depends on whether  $k$ -NNs are used for classification or regression [17]. Multilayer perceptions (MLP) is a class of feed-forward artificial



**TABLE 3** Model performance on 3-year survival prediction

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC
Logistic regression	0.684	0.702	0.670	0.686	0.683	0.741
Random forest	0.645	0.654	0.638	0.643	0.642	0.690
XGBoost	0.676	0.665	0.692	0.678	0.674	0.729
Decision tree	0.690	0.733	0.662	0.697	0.687	0.710
K-nearest neighbors	0.607	0.609	0.605	0.607	0.606	0.609
AdaBoost	0.675	0.675	0.675	0.675	0.675	0.736
Support vector machines	0.685	0.713	0.665	0.689	0.683	0.684
Multilayer perceptron	0.684	0.732	0.654	0.693	0.679	0.735

Abbreviation: AUROC, Area under the receiver operating characteristic.

**TABLE 4** Model performance on tumor metastasis prediction

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC
Logistic regression	0.800	0.540	0.830	0.769	0.772	0.804
Random forest	0.765	0.417	0.831	0.744	0.753	0.731
XGBoost	0.806	0.548	0.842	0.781	0.786	0.747
Decision tree	0.791	0.498	0.828	0.759	0.765	0.666
K-nearest neighbor	0.763	0.403	0.826	0.738	0.747	0.682
AdaBoost	0.797	0.518	0.846	0.777	0.784	0.799
Support vector machines	0.793	0.510	0.818	0.754	0.755	0.573
Multilayer perceptron	0.800	0.542	0.828	0.768	0.770	0.802

Abbreviation: AUROC, Area under the receiver operating characteristic.

**TABLE 5** The Logistic Regression performance on 3-year survival and tumor metastasis prediction for clear cell renal cell carcinoma

Performance on 3-year survival prediction		Performance on tumor metastasis prediction	
AUROC	0.710	AUROC	0.811
Accuracy	0.658	Accuracy	0.826
Sensitivity	0.674	Sensitivity	0.593
Specificity	0.649	Specificity	0.851
Precision	0.661	Precision	0.802
F1-score	0.653	F1-score	0.803

Abbreviation: AUROC, Area under the receiver operating characteristic.

neural networks (ANNs) that consist of at least three layers of nodes: the input layer, the hidden layer, and the output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP is trained using a supervised learning technique called back

propagation [18]. Its multilayer and nonlinear activation distinguish MLP from linear perceptions. It can distinguish nonlinearly separable data [18]. We compared the performance of these eight models, there were no significant difference in performance. It is hard to compare the models based on its algorithm, external validation should be conducted for model evaluation.

In addition, we used a single source of public data to ensure data completeness, which had limited generalizability. Although the variable selection was based on statistics and clinical experience, the overfitting was a common problem in machine learning models. We will subsequently optimize and balance the number of variables and the prediction results. In the future, this prediction models could be externally validated using hospital patient data. Furthermore, in real clinical scenarios, these clinical prediction models need to be transformed into useful tools after strong clinical validation, and they will be integrated into the hospital information system to provide timely decision support for future cancer care.

## 5 | CONCLUSION

The present study developed and compared eight machine learning models for predicting 3-year survival and metastasis of kidney cancer. The variable selection was based on both statistical significance and clinical experiences. In conclusion, Logistic Regression had the highest area under the receiver operating characteristic (AUROC) in both 3-year survival and metastasis prediction scenarios, and the prediction results potentially provide decision-making references to clinicians for early intervention and treatment.

### AUTHOR CONTRIBUTIONS

**Yuxiang Zhang:** Methodology; writing—original draft; writing—review and editing. **Na Hong:** Data curation; methodology; writing—review and editing. **Sida Huang:** Formal analysis. **Jie Wu:** Methodology. **Jianwei Gao:** Data curation; software. **Zheng Xu:** Software. **Fubo Zhang:** Software. **Shaohui Ma:** Writing—original draft. **Ye Liu:** Writing—original draft. **Peiyuan Sun:** Software. **Yanping Tang:** Supervision. **Chun Liu:** Data curation; software. **Jianzhong Shou:** Writing—review and editing. **Meng Chen:** Conceptualization; project administration; resources; supervision; writing—review and editing.

### ACKNOWLEDGEMENT

None.

### CONFLICT OF INTEREST

All authors declare that there is no conflict of interest except Professor Meng Chen, who is member of Cancer Innovation Editorial Board. To minimize bias, she was excluded from all editorial decision-making related to the acceptance of this study for publication.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in surveillance, epidemiology, and end results (SEER) database. These data were derived from the following resources available in the public domain: <https://seer.cancer.gov/data/>

### ETHICS STATEMENT

Not applicable.

### INFORMED CONSENT

Not applicable.

### ORCID

Yuxiang Zhang  <http://orcid.org/0000-0002-0393-922X>  
Na Hong  <http://orcid.org/0000-0001-6798-1761>

## REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7–30. <https://doi.org/10.3322/caac.21590>
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394–424. <https://doi.org/10.3322/caac.21492>
3. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F. et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* 2016;66(2): 115–32. <https://doi.org/10.3322/caac.21338>
4. National Cancer Institute. Surveillance, Epidemiology, and End Results program [Internet]. 2022 [cited 2022 Jun 6]. Available from: <http://www.seer.cancer.gov>
5. Scelo G, Larose TL. Epidemiology and risk factors for kidney cancer. *J Clin Oncol.* 2018;36(36):Jco2018791905. <https://doi.org/10.1200/JCO.2018.79.1905>
6. Vogelzang NJ, Stadler WM. Kidney cancer. *Lancet.* 1998;352(9141):1691–6. [https://doi.org/10.1016/S0140-6736\(98\)01041-1](https://doi.org/10.1016/S0140-6736(98)01041-1)
7. Thakur A, Jain SK. Kidney cancer: current progress in treatment. *World J Oncol.* 2011;2(4):158–65.
8. Meissner MA, McCormick BZ, Karam JA, Wood CG. Adjuvant therapy for advanced renal cell carcinoma. *Expert Rev Anticancer Ther.* 2018;18(7):663–71. <https://doi.org/10.1080/14737140.2018.1469980>
9. Byun SS, Heo TS, Choi JM, Jeong YS, Kim YS, Lee WK, et al. Deep learning based prediction of prognosis in nonmetastatic clear cell renal cell carcinoma. *Sci Rep.* 2021;11(1):1242. <https://doi.org/10.1038/s41598-020-80262-9>
10. Ji GW, Zhu FP, Xu Q, Wang K, Wu MY, Tang WW, et al. Machine-learning analysis of contrast-enhanced CT radiomics predicts recurrence of hepatocellular carcinoma after resection: a multi-institutional study. *EBioMedicine.* 2019;50: 156–65. <https://doi.org/10.1016/j.ebiom.2019.10.057>
11. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 2004;20:273–97. <https://doi.org/10.1007/BF00994018>
12. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *JAMA.* 2016;316(5):533–4. <https://doi.org/10.1001/jama.2016.7653>
13. Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. *Cent Eur J Oper Res.* 2018;26(1):135–59. <https://doi.org/10.1007/s10100-017-0479-6>
14. Tin Kam H. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition;* 1995.
15. XGBoost Documentation [Internet]. 2021 [cited 2022 Jun 6]. Available from: XGBoost Documentation—xgboost 1.5.0 documentation.
16. Kégl B. The return of AdaBoost MH: multi-class Hamming trees. 2013. arXiv:1312.6086. <https://doi.org/10.48550/arXiv.1312.6086>
17. Altman NS. An introduction to Kernel and nearest-neighbor nonparametric regression. *Am Stat.* 1992;46(3):175–85. <https://doi.org/10.1080/00031305.1992.10475879>
18. Van Der Malsburg C, editor. Frank Rosenblatt: principles of neurodynamics: perceptrons and the theory of brain mechanisms. *Brain theory.* Springer Berlin Heidelberg; 1986.

19. Evaluation of measurement data—Guide to the expression of uncertainty in measurement [Internet]. The JCGM member organizations; 2008 [updated 2008 Sep; cited 2022 Jun 6]. JCGM 100:2008 (GUM 1995 with minor corrections—Evaluation of measurement data ([bipm.org](http://bipm.org)).
20. Powers D, Ailab. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011;2:2229–3981.
21. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008;56(1):45–50. <https://doi.org/10.4103/0301-4738.37595>
22. Fawcett T. An introduction to ROC analysis. *Patt Recognit Lett*. 2006;27(8):861–74. <https://doi.org/10.1016/j.patrec.2005.10.010>
23. Lai Y, Tang F, Huang Y, He C, Chen C, Zhao J, et al. The tumour microenvironment and metabolism in renal cell carcinoma targeted or immune therapy. *J Cell Physiol*. 2021;236(3):1616–27. <https://doi.org/10.1002/jcp.29969>
24. Pavlovich CP, Schmidt LS. Searching for the hereditary causes of renal-cell carcinoma. *Nat Rev Cancer*. 2004;4(5):381–93. <https://doi.org/10.1038/nrc1364>
25. Rini B, Goddard A, Knezevic D, Maddala T, Zhou M, Aydin H, et al. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. *Lancet Oncol*. 2015;16(6):676–85. [https://doi.org/10.1016/S1470-2045\(15\)70167-1](https://doi.org/10.1016/S1470-2045(15)70167-1)
26. Brooks SA, Brannon AR, Parker JS, Fisher JC, Sen O, Kattan MW, et al. ClearCode34: a prognostic risk predictor for localized clear cell renal cell carcinoma. *Eur Urol*. 2014;66(1):77–84. <https://doi.org/10.1016/j.eururo.2014.02.035>

**How to cite this article:** Zhang Y, Hong N, Huang S, Wu J, Gao J, Xu Z, et al. Machine learning-based prognostic and metastasis models of kidney cancer. *Cancer Innovation*. 2022;1:124–134. <https://doi.org/10.1002/cai2.22>