

Prediction of Distant Metastasis of Renal Cell Carcinoma Based on Interpretable Machine Learning: A Multicenter Retrospective Study

Jinkai Dong^{1,*}, Minjie Duan^{2,3,*}, Xiaozhu Liu^{4,*}, Huan Li⁵, Yang Zhang⁶, Tingting Zhang⁷, Chengwei Fu¹, Jie Yu⁸, Weihe Hu⁹, Shengxian Peng^{9,*}

¹Senior Department of Urology, the Third Medical Center of PLA General Hospital, Beijing, People's Republic of China; ²Medical School of Chinese PLA, Beijing, People's Republic of China; ³Medical Innovation Research Department, Chinese PLA General Hospital, Beijing, People's Republic of China; ⁴Emergency and Critical Care Medical Center, Beijing Shijitan Hospital, Capital Medical University, Beijing, People's Republic of China; ⁵Chongqing College of Electronic Engineering, Chongqing, People's Republic of China; ⁶College of Medical Informatics, Chongqing Medical University, Chongqing, People's Republic of China; ⁷Department of Endocrinology, Fifth Medical Center of Chinese PLA Hospital, Beijing, People's Republic of China; ⁸Department of Medical Imaging, The Affiliated Taian City Central Hospital of Qingdao University, Taian, People's Republic of China; ⁹Scientific Research Department, First People's Hospital of Zigong City, Zigong, People's Republic of China

*These authors contributed equally to this work

Correspondence: Shengxian Peng, Scientific Research Department, First People's Hospital of Zigong City, No. 42, 1st Street, Shangyihao, Ziliujing District, Zigong, Sichuan, 643000, People's Republic of China, Email 13258280319@163.com

Background: The traditional tool for predicting distant metastasis in renal cell carcinoma (RCC) is still insufficient. We aimed to establish an interpretable machine learning model for predicting distant metastasis in RCC patients.

Methods: We involved a population-based cohort of 121433 patients (mean age = 63 years; 63.58% men) diagnosed with RCC between 2004 and 2015 from the Surveillance, Epidemiology, and End Results (SEER) database. The lightGBM algorithm was used to develop prediction model and assessed by the area under the receiver-operating characteristic curve (AUC), accuracy, sensitivity, and specificity. The LightGBM model was then externally validated in 36395 RCC patients enrolled from the SEER database between 2016 and 2018. Shapley Additive exPlanations (SHAP) method was applied to provide insights into the model's outcome or prediction.

Results: Of 121433 patients involved in the study cohort, 10730 (8.84%) had distant metastasis. The LightGBM model showed good performance in the internal validation set (AUC: 0.955, 95% CI: 0.951–0.959) and temporal external validation sets (0.963, 95% CI: 0.959–0.967; 0.961, 95% CI: 0.954–0.966). Performance for the prediction model was also well performed in different sub-cohort stratified by age, gender, and ethnicity. The calibration curve indicated that the predicted values are highly consistent with the actual observed values. SHAP plots demonstrated that chemotherapy was the most vital variable for prediction of distant metastasis of RCC patients.

Conclusion: We developed an interpretable machine learning model that is capable of accurately predicting the risk of distant metastasis of RCC patients. The presented model could help identify high-risk patients who require additional treatment strategies and follow-up regimens.

Keywords: distant metastasis, machine learning, renal cell carcinoma, prediction, interpretable

Introduction

Renal cell carcinoma (RCC) is the most prevalent pathological subtype of kidney neoplasm and the 6th most common genitourinary malignancy in males, trailing only prostate and bladder cancer.¹ Simultaneously, it is the deadliest malignant urological tumour, accounting for around 2% of all cancer fatalities globally.² RCC consists of heterogeneous tumor cells, with the most common histological subtypes being clear cell renal cell carcinoma (ccRCC), papillary renal cell carcinoma (pRCC), and chromophobe renal cell carcinoma (crRCC).^{3,4} The prognosis of RCC patients with various

histological subtypes varies, and it has been demonstrated that patients with ccRCC have a poorer prognosis than patients with other histological subtypes (pRCC and crRCC) even after radical nephrectomy.^{5,6}

RCC with distant metastasis is prevalent, especially in individuals with high-grade, microvascular invasion, and pathological sarcomatoid, papillary, or chromophobe renal cell carcinoma. Regional lymph nodes and the lungs are the most prevalent probable sites of metastasis for RCC, followed by the brain, bones, ovaries, and soft tissues.^{7–13} The distant metastasis was highly associated with the primary tumor's histological grade, T staging, gene expression, and functional features.¹⁴ According to prior studies, individuals with distant metastasis had a worse prognosis than those who did not.¹⁵ The 5-year survival rates of RCC in TNM stages I and II were as high as 81% and 74%, respectively, while the 5-year survival rates of RCC in stages III and IV were only 53% and 10%. The emergence of distant metastasis is the primary cause of the poor survival rate.^{16–18} Additionally, it has been demonstrated that individuals with metastatic RCC may benefit from cancer-targeted surgical surgery for a better prognosis.^{19,20} As a result, it is vital to identify the clinicopathological risk factors that promote RCC distant metastasis.

In recent years, there have been some clinical diagnostic tools built using sophisticated machine learning algorithms. Compared to traditional statistical regression, machine learning can explore larger and multi-dimensional data and better describe complicated correlations between risk factors and outcomes.^{21,22} N. P. Singh et al created a machine learning method to discover biomarkers and built classifiers to differentiate between early and late stages of pRCC based on gene expression profiles.²³ S. T. Chen et al created a machine learning-based pathomics signature (MLPS) for patients with ccRCC, which can be used as a novel prognostic marker.²⁴ A machine learning method was created by H. Kim et al to forecast the likelihood of RCC recrudescence between 5 and 10 years after surgery.²⁵ However, a promising machine learning-based model has not been introduced for prediction of distant metastases in RCC patients. Hence, considering the poor prognosis of patients with metastatic RCC (mRCC), we sought to develop a machine learning model to predict the risk of distant metastases in RCC patients, and visualize the clinical features and their interactions that affect distant metastases in RCC patients using Shapley Additive exPlanations framework, which has potential clinical utility.

This study was reported according to the recommendations of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.^{26,27}

Methods

Study Population and Data Source

The study cohort was extracted from the Surveillance, Epidemiology, and End Results (SEER) database from the National Cancer Institute. Basic demographic information, clinical characteristics, and survival status of patients are all publicly available through the SEER database (<https://seer.cancer.gov/>, ID: 15899-Nov2020). The inclusion criteria are as follows: patients pathologically diagnosed with histologically confirmed RCC (site code 64, 65, morphology code 8310, 8260, 8312, 8317, 8255, 8318) between Jan 1, 2004, and Dec 31, 2015; the exclusion criteria are as follows: (1) unknown surgery type; (2) unknown tumor size; (3) unknown T stage; (4) survival time <1 month; (5) unknown N stage; (6) bilateral renal masses. Data of RCC patients meeting the same criteria from 2016 to 2018 from the SEER database were extracted for temporal external validation. Due to the tumor staging of patients in 2016–2017 was mainly based on the 7th edition of the American Joint Committee on Cancer (AJCC) TNM Staging Manual, while the tumor staging of patients in 2018 was based on the 8th edition of AJCC TNM Staging Manual, the temporal validation set was divided into two sets: set 1 (2016–2017) and set 2 (2018). Since the data is public and anonymous and open access policy of the SEER database, ethical review and patients' informed consent were waived.

The primary outcome of this study was distant metastasis in RCC patients. Distant metastasis of RCC refers to the presence of metastatic lesions in other sites besides primary renal lesions.

Data Extraction

Demographic and clinical information including age at diagnosis, gender, race, laterality, T stage, N stage, histological type, Fuhrman grade, surgery, chemotherapy, radiotherapy, tumor size, Lymph Node Surgery, and marital status were extracted from the SEER database. The patient's race includes black, white, and other races (American Indian/AK

Native, Asian/Pacific Islander). Fuhrman grades I, II, III, and IV represent well-differentiated, moderately differentiated, poorly differentiated, and undifferentiated, respectively. The surgical methods are divided into four groups according to the SEER Kidney Surgery Codes 2018: non-surgical group (code 0), local excision (code 10–27), partial nephrectomy (PN, code 30), and radical nephrectomy (RN, code 40–80).

Model Development and Assessment

The study cohort was split into a training cohort (70%) and an internal validation cohort (30%) using stratified random sampling. First, this study compared the performance of 15 machine learning algorithms with default hyperparameters to screen the best algorithms for further prediction models. The screening was carried out with the PyCaret package (version 2.3.3), an open-source and low-code ML library in Python. Due to the low proportion of positive case (8.84%), we applied several adaptive sampling methods, including Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling, Tomek links, and Random Undersampling, to handle the class imbalance problem and improve the model performance. Bayes optimization method with ten-fold cross-validation was applied to choose the best hyperparameters setting. The hyperparameter tuning process was implemented using the bayes_opt python package (version 1.2.0). We mainly used the area under the receiver operating characteristic curve (AUC) to assess the discrimination of prediction models. The accuracy, sensitivity, specificity, and F1_score were also calculated to comprehensively assess models. Meanwhile, we also used the area under the precision–recall curve (AUPRC), which can provide a robust metric for unbalanced datasets. Prediction performance was also measured by calibration curve, which evaluates how well-predicted probability matches the observed probability. Performance was assessed in different sub-groups stratified by age (20–40, 40–60, >60), gender (male, female), and ethnicity (Black, White, other). Bootstrapping was used to calculate 95% confidence intervals (CIs) of performance metrics.

To alleviate the black-box nature of ML and help clinicians grasp the results offered by models, Shapley Additive exPlanations (SHAP) was applied to interpret the prediction results of the model using python shap package (version 0.39.0).

Statistical Analysis

In this study, the baseline characteristics of RCC patients were compared between the groups with distant metastasis and those without it. The chi-square test was used to compare categorical variables that were presented as frequency and percentages. Continuous variables were reported as median with the first quartile (Q1) and the third quartile (Q3) and compared using Mann–Whitney *U*-test. A two-sided $P < 0.05$ was judged statistically significant. Statistical analyses were conducted with an open-source Scipy python package (version 1.7.1).

Results

Baseline Characteristics of the Participants

The flowchart for the cohort study was shown in [Figure 1](#). A total of 121,433 patients were involved with histologically confirmed RCC from SEER database between Jan 1, 2004, and Dec 31, 2015. These patients were randomly divided into a training set ($n = 85003$) and a validation set ($n = 36430$). The detail demographic clinicopathological information of the patients was shown in [Table 1](#). Of these patients, the mean age was 63 years, most patients were White (82.09% 99684 of 121433), fewer patients were Black and other race (11.49% and 6.42%). A total of 77,213 (63.58%) were male, 74498 (61.35%) were married, 70748 (58.26%) with the histopathological type of renal clear cell carcinoma, 116146 (95.65%) were N0, 84,389 (69.49%) were T1 stage, 109364 (90.06%) were LN. Sur and the average tumor diameter were 41.0mm. Most patients underwent surgery, 67132 (55.28%) patients underwent radical nephrectomy, 114341 (94.16%) patients did not receive chemotherapy, and 117812 (97.02%) patients did not receive radiotherapy. There is no significant difference was found in the characteristics of patients between the training set and the validation set. Characteristics of patients in the external validation sets were shown in [Tables S1](#) and [S2](#).

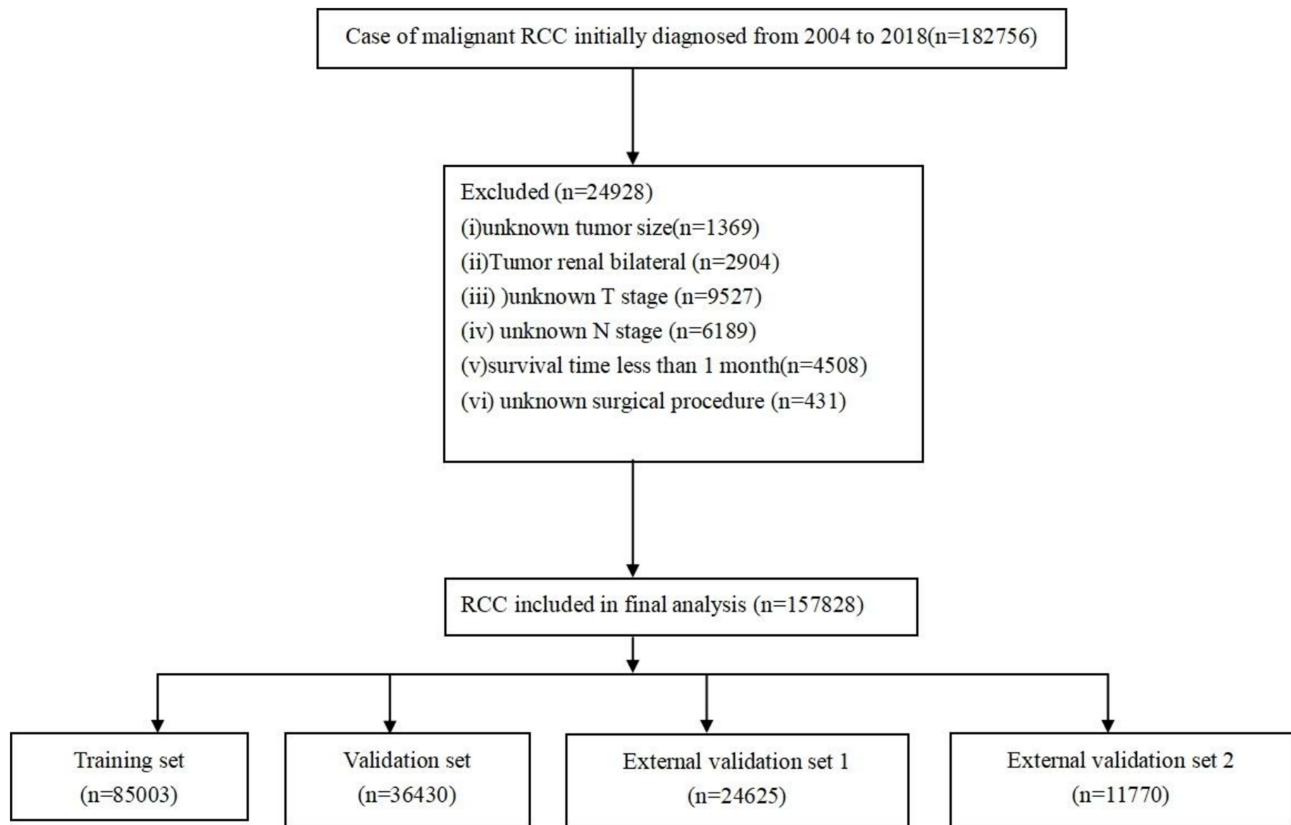


Figure 1 Study design flowchart of specific patient screening process.

Model Performance

These patients were randomly divided into a training set (n = 85003) and an internal validation set (n = 36430). The results of 15 algorithms were shown in [Table S3](#). The Light Gradient Boosting Machine (LightGBM) was selected to further develop the prediction models according to accuracy and AUC. Owing to the extreme class imbalance that exists in our dataset, we selected random undersampling to address the class imbalance problem ([Table S4](#)). After the process of hyperparameter tuning, the LightGBM model showed great performance (AUC: 0.955, 95% CI: 0.951–0.959) based on the independent internal validation set. The ROC curve of the LightGBM model was shown in [Figure 2A](#). The accuracy,

Table 1 Demographics and Clinical Characteristics for the Training Set and Validation Set

	Overall (n=121433)	Training Set (n=85003)	Validation Set (n=36430)	P
Age (median [IQR])	63.00 [54.00, 71.00]	63.00 [54.00, 71.00]	63.00 [54.00, 72.00]	0.4244
Race (%)				0.5329
White	99684 (82.09)	69,845 (82.17)	29,839 (81.91)	
Black	13953 (11.49)	9715 (11.43)	4238 (11.63)	
Other	7796 (6.42)	5443 (6.40)	2353 (6.46)	
Sex (%)				0.7201
Male	77213 (63.58)	54,021 (63.55)	23,192 (63.66)	
Female	44220 (36.42)	30,982 (36.45)	13,238 (36.34)	
Marital (%)				0.8474
No	46935 (38.65)	32,839 (38.63)	14,096 (38.69)	
Married	74498 (61.35)	52,164 (61.37)	22,334 (61.31)	

(Continued)

Table 1 (Continued).

	Overall (n=121433)	Training Set (n=85003)	Validation Set (n=36430)	P
Histologic type (%)				0.0747
Clear cell	70748 (58.26)	49,652 (58.41)	21,096 (57.91)	
Papillary	15316 (12.61)	10,590 (12.46)	4726 (12.97)	
Chromophobe	6647 (5.47)	4676 (5.50)	1971 (5.41)	
Not classified	28722 (23.65)	20,085 (23.63)	8637 (23.71)	
Grade (%)				0.9353
I	12167 (10.02)	8544 (10.05)	3623 (9.95)	
II	49646 (40.88)	34,772 (40.91)	14,874 (40.83)	
III	26712 (22.00)	18,708 (22.01)	8004 (21.97)	
IV	5748 (4.73)	4018 (4.73)	1730 (4.75)	
Unknown	27160 (22.37)	18,961 (22.31)	8199 (22.51)	
Laterality (%)				0.0986
Left	59811 (49.25)	42,000 (49.41)	17,811 (48.89)	
Right	61622 (50.75)	43,003 (50.59)	18,619 (51.11)	
T (%)				0.1708
T1a	56761 (46.74)	39,725 (46.73)	17,036 (46.76)	
T1b	27628 (22.75)	19,225 (22.62)	8403 (23.07)	
T2	13656 (11.25)	9654 (11.36)	4002 (10.99)	
T3	23386 (19.26)	16,397 (19.29)	6989 (19.18)	
T4	2 (0.00)	2 (0.00)	0 (0.00)	
N (%)				0.1636
N0	116146 (95.65)	81,348 (95.70)	34,798 (95.52)	
N1	5287 (4.35)	3655 (4.30)	1632 (4.48)	
LN.Sur (%)				0.7361
0	109,364 (90.06)	76,527 (90.03)	32,837 (90.14)	
I-3	7000 (5.76)	4903 (5.77)	2097 (5.76)	
≥4	5069 (4.17)	3573 (4.20)	1496 (4.11)	
Surgery (%)				0.3914
No	13899 (11.45)	9727 (11.44)	4172 (11.45)	
Local tumor excision	6436 (5.30)	4527 (5.33)	1909 (5.24)	
Partial nephrectomy	33966 (27.97)	23,658 (27.83)	10,308 (28.30)	
Radical nephrectomy	67132 (55.28)	47,091 (55.40)	20,041 (55.01)	
Radiation (%)				0.5029
No/Unknown	117812 (97.02)	82,487 (97.04)	35,325 (96.97)	
Yes	3621 (2.98)	2516 (2.96)	1105 (3.03)	
Chemotherapy (%)				0.7306
No/Unknown	114341 (94.16)	80,052 (94.18)	34,289 (94.12)	
Yes	7092 (5.84)	4951 (5.82)	2141 (5.88)	
Tumor size (median [IQR])	41.00[27.00, 66.00]	41.00[27.00, 67.00]	41.00 [27.00, 65.00]	0.5318
Mets at dx (%)				1
No/Unknown	110703 (91.16)	77,492 (91.16)	33,211 (91.16)	
Yes	10730 (8.84)	7511 (8.84)	3219 (8.84)	

sensitivity, specificity, and F1_score were listed in Table 2. The areas under the precision-recall curve (AUPRCs) are more informative and was 0.768 (Figure 2B). As shown in Figure 2C, the calibration plot showed that the lightGBM presented excellent goodness of fit.

Furthermore, we validated model performance in different age, gender, and ethnic sub-groups. Patients younger than 60 had a marginally higher AUC than those older than 60 (Figure 3A), with AUCs of 0.961, 0.963, and 0.949, respectively. The AUCs of the distant metastasis were 0.954 and 0.956 for men and women group, respectively

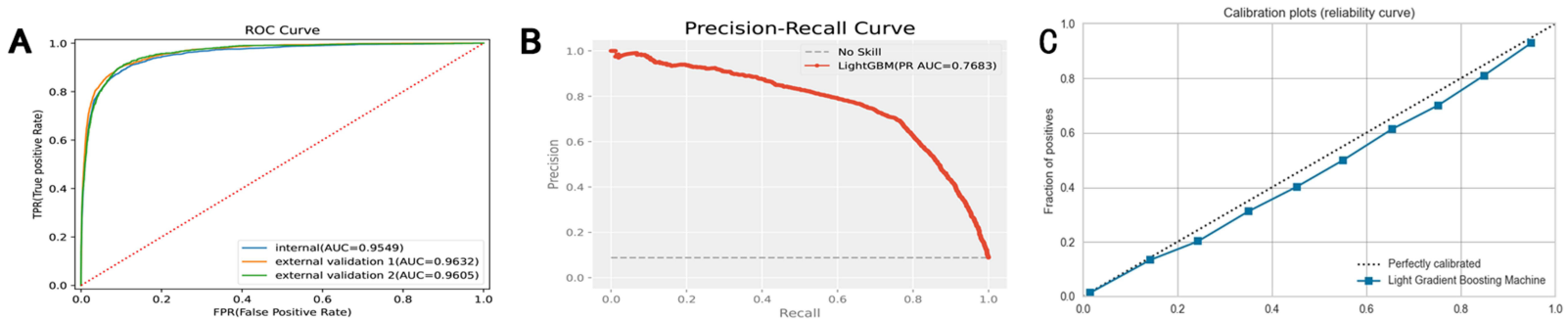


Figure 2 The performance of LightGBM model. The ROC curves of the model in three validation sets (**A**). The precision–recall curve of the model (**B**). Calibration curve of the model (**C**).

Table 2 Internal Validation and External Validation Performance for the Prediction of Distant Metastasis in Patients with RCC

	AUC	Accuracy	Sensitivity	Specificity	F1_score
Internal validation set	0.9549 (0.9510–0.9586)	0.8993 (0.8962–0.9023)	0.8823 (0.8712–0.8933)	0.9009 (0.8977–0.9041)	0.6076 (0.5959–0.6190)
External validation set 1	0.9632 (0.9595–0.9667)	0.8970 (0.8932–0.9008)	0.9058 (0.8940–0.9171)	0.8961 (0.8921–0.9001)	0.6396 (0.6262–0.6525)
External validation set 2	0.9605 (0.9544–0.9660)	0.8979 (0.8924–0.9033)	0.9055 (0.8872–0.9228)	0.8972 (0.8914–0.9029)	0.6072 (0.5860–0.6278)

(Figure 3B). The AUCs of the distant metastasis were 0.956, 0.946 and 0.956 for white, black and other race, respectively (Figure 3C).

Temporal External Validation

In the temporal validation sets 1 and 2, the incidence of distant metastasis is 10.1% and 8.7%, respectively. The ROC curves for the external validation sets 1 and 2 were showed in Figure 2A. The discrimination for predicting distant metastasis was consistently high in external validation set 1 (AUC: 0.963, 95% CI: 0.960–0.967) and external validation set 2 (AUC: 0.961, 0.954–0.966). The accuracy, sensitivity, specificity, and F1_score analyses were shown in Table 2.

Model Interpretation

We provided insights into the model's prediction outcome at the global and individual level using the Tree-Explainer class imported from the SHAP package based on the independent internal validation set. As shown in Figure 4, chemotherapy, surgery, tumor size, T stage, and radiation were the top 5 ranked variables that were correlated with distant metastasis among patients with RCC, followed by grade, histologic type, N stage, age, LN. Sur, race, sex, laterality, and marital, positive and negative shapley values indicate the positive and negative impact on the prediction outcome.

Four samples were demonstrated as the SHAP force plot in Figure 5. For example, in Figure 5A, this is a 63-year-old man with a chemotherapy history and 87mm of tumor size. Factors identified by the model that increased the risk of distant metastasis were chemotherapy = yes, surgery = no, N = N1. The elevated risk is offset by the radiation = no. The ML model's predicted outcome was the occurrence of distant metastasis, and the actual outcome was also the occurrence of distant metastasis (True Positive).

In Figure 5B, this is an 83-year-old man with a tumor size of 46mm and no history of chemotherapy and radiation therapy. The ML model counteracts the increased risk of the patient because the patient had not received chemotherapy and radiotherapy, the tumor size was 46mm, and the surgical procedure was a radical nephrectomy. The ML model incorrectly predicted the outcome as free from distant metastasis for this patient, whereas the actual outcome was distant metastasis (False Negative).

In Figure 5C, this is a 74-year-old woman with a tumor size of 31mm and no history of chemotherapy and radiation therapy. Factors decreasing the risk included tumor size = 31.0, chemotherapy = no, T = T1a, radiation = no, and N = N0. The predicted outcome of the ML model was that the patient had non-distant metastases, while the actual outcome was also non-distant metastases (True Negative).

In Figure 5D, this is a 41-year-old unmarried man with a tumor size of 60mm and no history of chemotherapy and radiation therapy. The patient's risk for distant metastasis increased because of larger tumor size, higher tumor grade, and histological subtypes (ccRCC). Factors offsetting the risk were no chemotherapy, no radiation, and undergoing a radical nephrectomy. The outcome predicted by the ML model was the patient's distant metastasis, while the actual outcome was non-distant metastasis (False Positive).

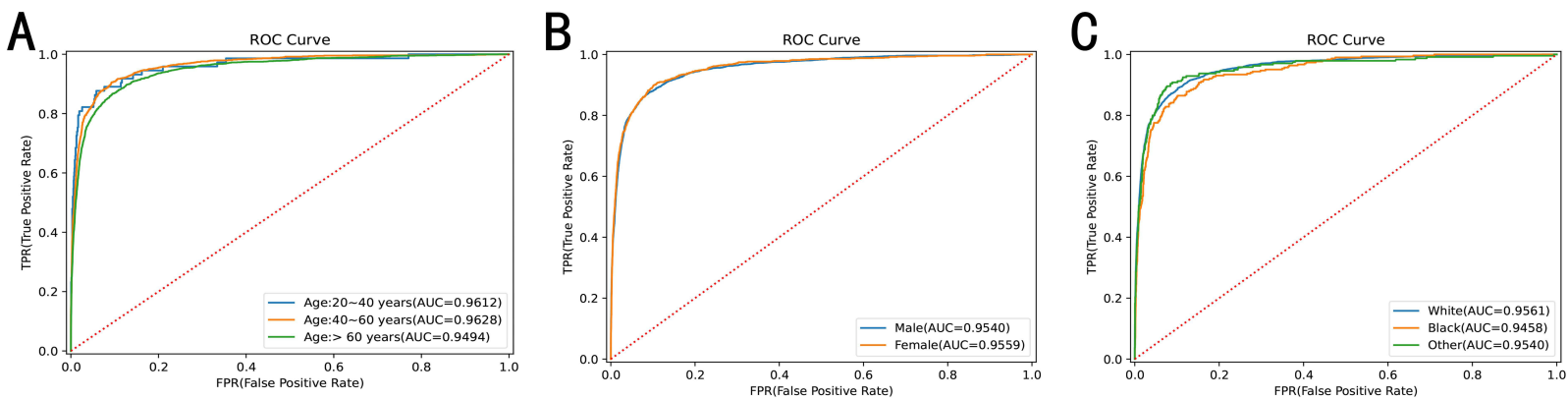


Figure 3 The ROC curve for different sub-groups. The predictive performance of the LightGBM model for different age sub-cohorts (A), genders (B), and races (C) in the independent validation set.

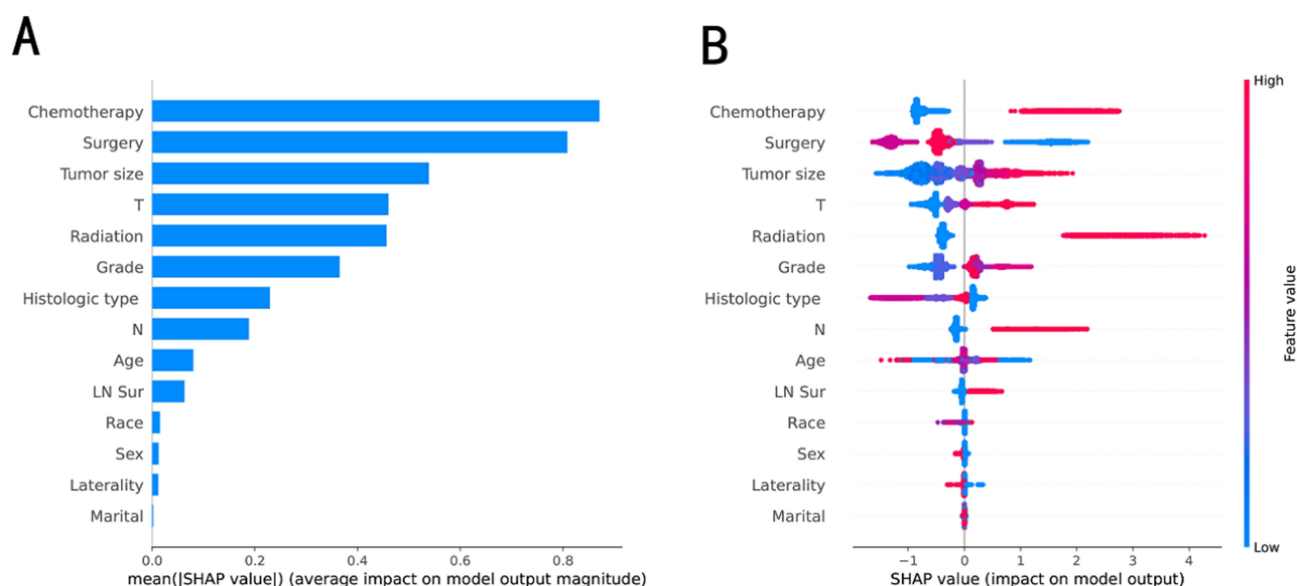


Figure 4 SHAP summary plot for the 14 clinical features contributing to our ML model's prediction for distant metastasis (**A**). The distribution of the impacts of each feature on the model output (**B**). The colors represent the feature values for numeric features: red for larger values and blue for smaller ones. The line is made of individual dots representing each RCC patient, and the thickness of the line is determined by the number of examples at a given value. A negative SHAP value (extending to the left) indicates a reduced probability, while a positive one (extending to the right) indicates an increased probability. LN sur, Lymph Node Surgery.

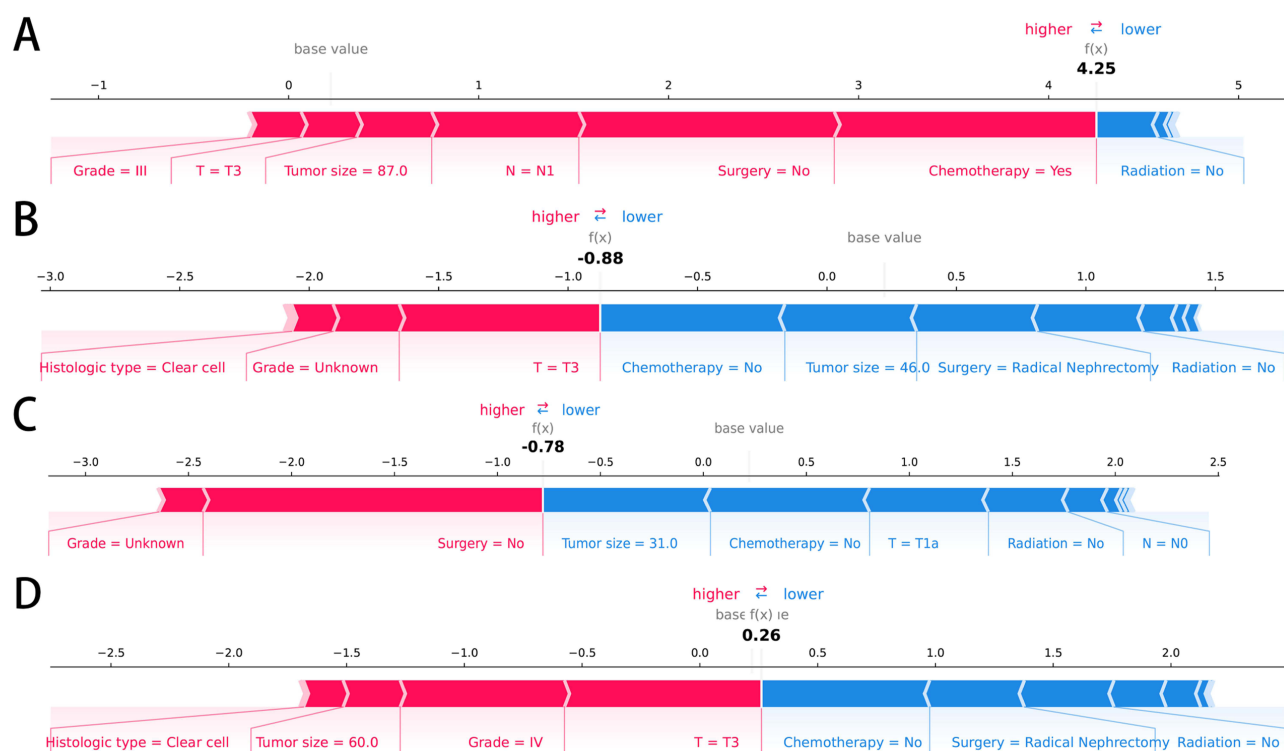


Figure 5 SHAP force plot for four samples from the internal validation set. The bars in red and blue represent risk factors and protective factors, respectively; longer bars represent greater feature importance. (**A**) The model correctly predicted the presence of distant metastasis, indicating a True Positive. (**B**) The model failed to predict the presence of distant metastasis, indicating a False Negative. (**C**) The model correctly predicted the absence of distant metastasis, indicating a True Negative. (**D**) The model incorrectly predicted the presence of distant metastasis, indicating a False Positive.

Variable Interactions

The interactions between variables can be analyzed by SHAP dependence plot. The top five variables were shown in [Figure S1](#). For example, the size of the tumor was shown to affect the probability of distant metastasis in RCC patients differently for patients who underwent a type of surgery. With radical nephrectomy and partial nephrectomy, there is

a more pronounced trend that smaller tumor size at discharge contributed to a lower risk of distant metastasis, while larger tumor size contributed to an increased risk of distant metastasis (SHAP values from -1.5 to $+0.5$ – 1.5). With local excision and no surgery, larger tumor size overall is observed, and the positive trend is more remarkable. This may highlight the importance of considering a variable such as a tumor size in a more complete clinical setting, such as one that includes tumor size, chemotherapy, type of surgery, grade, histologic type, and clinical reasoning (e.g. Patients with smaller tumors who have undergone radical nephrectomy and partial nephrectomy are not likely to be discharged with a significant risk of distant metastasis (Figure S1C). In addition, a similar finding is observed in Figure S1B for distant metastasis prediction, although clinical reasoning is unlikely to work compared to more purely physiologic phenomena: larger tumor overall are observed in patients who underwent local excision, and the relationship between tumor size and distant metastasis is not observed to be as linear for local excision (For certain types of surgery, high and low SHAP values were observed relatively consistently concerning tumor size.).

Discussion

RCC is a harmful malignant tumor of urinary system. The occurrence of distant metastatic disease is a routine marker of its advanced stage, which usually suggests that the prognosis of RCC patients is unfavorable.^{19,28} A prior study discovered that up to 18%–30% of RCC patients had systemic metastases at the time of their first diagnosis.²⁹ It is well established that RCC patients with distant metastasis highly influence the prognosis of patients.³⁰ Previous studies have demonstrated that radionic features for kidney cancer could be utilised as prognostic or predictive markers.

Meanwhile, it has a good ability to predict histological subtype, pathological grade, and treatment effects.^{31,32} U. Capitanio et al²⁸ found that computed tomography is still the current the first-line approach for detecting probable metastatic disease due to its high staging accuracy.

However, CT examination still has limitations, especially in evaluating the status of lymph nodes.²¹ Consequently, it is crucial for timely as well as accurate prediction of the risk of distant metastasis to potentially guide an optimal treatment decision.

Although conventional statistical analyses such as multivariate logistic regression analyses can predict the probability of RCC distant metastasis, their ability to handle large-scale and multi-dimensional data is limited. In this study, we established a machine learning prediction model to predict the risk of RCC distant metastasis by selecting 14 variables that significantly affect distant metastasis based on data collected from the SEER database. The LightGBM model showed good performance, indicating that this model may have a beneficial impact on decision-making in clinical practice. In addition, the AUC of the model varied from 0.949 to 0.963 for models with distant metastasis at different ages and varied from 0.955 to 0.956 at different race group. The model showed comparable performance between female and male group, indicating the robustness of the present model. We found that chemotherapy, surgery, tumor size, T stage, radiation, grade, histologic type, and N stage were significant features for predicting distant metastasis in RCC patients through SHapley Additive exPlanations framework.

The immune score system, functioning as an assessment approach, has demonstrated significant prognostic value in the treatment of renal cell carcinoma, particularly in forecasting survival and potential adjuvant treatment responses.³³ A high immune score is conspicuously associated with a poor prognosis, which might facilitate clinicians in more effectively evaluating treatment options and prognosis. Based on these discoveries, further research and validation are requisite for the effective integration of immune scoring into the treatment and management of RCC patients. Additionally, by incorporating immune score classification into the classic TNM staging system, we are capable of developing more intensive follow-up protocols for high-risk patients. If we can conduct a more in-depth analysis of the most relevant clinical parameters for each patient to transform it into an effective clinical tool for personalized assessment, it can be employed to predict the incidence of distant metastasis, potentially enhancing the survival outcomes of RCC patients.

Compared with previously nomogram model for distant metastasis prediction in RCC patients, the lightGBM model in our study showed better performance in discrimination and calibration.³⁴ In addition, the good generalization of our model has been proved by the temporal external validation in our study.

With the rapid development of machine learning, predictive tools that can be used to analyze and infer relationships between clinical characteristics and patient outcomes have spawned a wide range of clinical applications.^{35,36} Given the weaknesses of traditional statistical methods in handling large data containing highly correlated covariates, we employed an advanced machine learning algorithm called light gradient boosting machine (lightGBM) to develop a prediction model and identify the important clinical features for predicting the occurrence of distant metastasis in RCC patients. LightGBM is an ensemble algorithm developed by Microsoft, which provides an efficient operation for the gradient boosting algorithm. The LightGBM algorithm is characterized by lower memory usage, higher speed and efficiency, and more compatibility with large datasets than other boosting algorithms.³⁷ In addition, compared to other ML methods, LightGBM achieved state-of-the-art performance, especially in structured data.³⁸ ROC and AUC are popular and powerful measure for evaluating the performance of machine learning models with binary classifiers. However, special caution is needed when working with class unbalanced data. Precision–recall curve has been suggested as substitutes for ROC but are used infrequently. Previous analysis of the literature found that most studies of algorithms using unbalanced datasets used ROC as their primary method of performance evaluation.³⁹ This study showed that PR curve represented the sensitivity to unbalanced datasets with clear visual signals and allowed for accurate and intuitive explanations of actual model performance.

Recently, machine learning has emerged as a promising tool for identifying adverse outcomes that may be missed by clinicians. However, the nature of machine learning, black box, has become a major obstacle to its widespread application in clinical practice. Interpretability can be defined as the degree to which clinicians can understand the mechanism for the machine learning model's prediction.^{3,25} In order to illustrate how these clinical features influence the distant metastasis and made the model “explainable”, we applied SHAP framework to show whether each feature contributed positively or negatively to the patient's outcome. In this study, our analysis provides more direct conclusions and a way to visualize the relationships that produce the predicted results. We found that chemotherapy, surgery, tumor size, T stage, radiation, grade, histologic type, and N stage, were significant features for predicting distant metastasis. Especially, we found that poorly differentiated and undifferentiated patients had a greater risk of distant metastasis than well-differentiated patients, which supports previously reported findings that RCC differentiation was an independent risk factor for distant metastasis in RCC survivors.⁴⁰

The SHAP plot demonstrated that surgery is the second important feature of distant metastasis for tumor metastasis. Patients who undergo surgery have a lower risk of distant metastases than those who do not. This may be the outcome of efficient surgical removal of the tumor, which lowers the possibility of tumour spread.³⁴ For distant metastasis of RCC, we discovered that tumour size is the third crucial factor. The chance of distant metastasis increases with tumour size. Our findings are consistent with earlier research.⁴¹

This study has various limitations that should be mentioned. First, our analysis was based on the SEER database and was not independently evaluated on a separate dataset or at our hospital. Second, because we used retrospective data from the SEER database, our study may have had some inherent bias. To validate our prediction model in the general population, prospective clinical studies with a larger sample size are necessary. Third, while our approach is a clear explanation of the full prediction process, it does not explain the effect of the variables throughout the entire range of potential values, instead highlighting which variables are critical. Fourth, we were unable to obtain some biomarkers from the SEER database, such as plasma osteopontin (OPN), miR-646, and epigenetic markers. To this end, a prospective study is planned. To summarise, as with all MLs that explore cause-and-effect correlations, this is a hypothesis-generating activity that requires strict validation, independent study on promising elements, and finally application at the discretion of patients and clinicians. We expect that putting greater focus on intelligence augmentation, decision support, and interpretability will result in a more sophisticated and effective adoption of ML as a supplementary tool in a comprehensive approach to patient care and research.

Conclusion

In this study, we applied a novel ML algorithm called LightGBM to predict the risk of distant metastasis and the model achieved great performance. SHapley Additive exPlanations framework was used to show the quantitative attributes

linked to the probability of distant metastasis in RCC patients and visualize how these attributes continuously affect the distant metastasis. The presented prediction model has the potential to be used to support clinical decision-making and treatment strategies.

Data Sharing Statement

The data analyzed in this study are available at <https://seer.Cancer.gov/>.

Ethics Statement

The Medical Ethics Committee of Zi Gong First People's Hospital has reviewed and approved this work (the ethics review number is No.39 of 2022). The data of this study originated from the SEER database. The data of the patients are publicly available anonymous data, and this study is of a retrospective nature. Therefore, obtaining consent from the informed patients was not necessary.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

Jinkai Dong, Minjie Duan, and Xiaozhu Liu are co-first authors for this study. Shengxian Peng is the senior author of this study. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Gonzalez Leon T, Morera Perez M. Renal cancer in the elderly. *Curr Urol Rep*. 2016;17(1). doi:10.1007/s11934-015-0562-2
- 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. doi:10.3322/caac.21492
- Cheville JC, Lohse CM, Zincke H, Weaver AL, L M. Blute LM Comparisons of outcome and prognostic features among histologic subtypes of renal cell carcinoma. *Am J Surg Pathol*. 2003;27(5):612–624. doi:10.1097/00000478-200305000-00005
- Gudbjartsson T, Hardarson S, Petursdottir V, Thoroddsen A, Magnusson J, Einarsson GV. Histological subtyping and nuclear grading of renal cell carcinoma and their implications for survival: a retrospective nation-wide study of 629 patients. *Europ Urol*. 2005;48(4):593–600. doi:10.1016/j.eururo.2005.04.016
- Liu LF, Qi L, Li YL, et al. Retroperitoneoscopic partial nephrectomy for moderately complex ventral hilar tumors: surgical technique and trifecta outcomes from a single institution in China. *J Laparoendosc Adv Surg Tech*. 2017;27(8):812–817. doi:10.1089/lap.2016.0194
- Yi XP, Xiao Q, Zeng FY, et al. Chen: computed tomography radiomics for predicting pathological grade of renal cell carcinoma. *Front Oncol*. 2021;10:570396. doi:10.3389/fonc.2020.570396
- Kavolius JP, Mastorakos DP, Pavlovich C, Russo P, Burt ME, Brady MS. Resection of metastatic renal cell carcinoma. *J Clin Oncol*. 1998;16(6):2261–2266. doi:10.1200/jco.1998.16.6.2261
- Hoffmann NE, Gillett MD, Cheville JC, Lohse CM, Leibovich BC, L M. Blute LM: Differences in organ system of distant metastasis by renal cell carcinoma subtype. *J Urol*. 2008;179(2):474–477. doi:10.1016/j.juro.2007.09.036
- Nguyen MM, Gill IS. Effect of renal cancer size on the prevalence of metastasis at diagnosis and mortality. *J Urol*. 2009;181(3):1020–1027. doi:10.1016/j.juro.2008.11.023
- Ku JH, Moon KC, Kwak C, Kim HH. Metachronous metastatic potential of small renal cell carcinoma: dependence on tumor size. *Urology*. 2009;74(6):1271–1275. doi:10.1016/j.urology.2009.04.072
- Kume H, Suzuki M, Fujimura T, et al. Distant metastasis of renal cell carcinoma with a diameter of 3 cm or less-which is aggressive cancer? *J Urol*. 2010;184(1):64–68. doi:10.1016/j.juro.2010.03.019
- Hellenthal NJ, Stewart GS, Cambio AJ, DeLair SM. Renal cell carcinoma metastatic to gallbladder: a survival advantage to simultaneous nephrectomy and cholecystectomy. *Int Urol Nephrol*. 2007;39(2):377–379. doi:10.1007/s11255-006-9057-x
- Iesalnieks I, Trupka A, Raab M, et al. Renal cell carcinoma metastases to the thyroid gland - 8 cases reported. *Thyroid*. 2007;17(1):49–52. doi:10.1089/thy.2006.0176
- Yu J, Xu Q, Huang DY, et al. Shi: prognostic aspects of dynamic contrast-enhanced magnetic resonance imaging in synchronous distant metastatic rectal cancer. *Eur Radiol*. 2017;27(5):1840–1847. doi:10.1007/s00330-016-4532-y
- Bruno JJ, Snyder ME, Motzer RJ, Russo P. Renal cell carcinoma local recurrences: impact of surgical treatment and concomitant metastasis on survival. *BJU Int*. 2006;97(5):933–938. doi:10.1111/j.1464-410X.2006.06076.x

16. Porta C, Cosmai L, Leibovich BC, Powles T, Gallieni M, Bex A. The adjuvant treatment of kidney cancer: a multidisciplinary outlook. *Nat Rev Nephrol.* **2019**;15(7):423–433. doi:10.1038/s41581-019-0131-x
17. Yoshiyama A, Morii T, Susa M, et al. Preoperative evaluation of renal cell carcinoma patients with bone metastases on risks for blood loss, performance status and lethal event. *J Orthop Sci.* **2017**;22(5):924–930. doi:10.1016/j.jos.2017.07.006
18. Abdel-Rahman O. Clinical correlates and prognostic value of different metastatic sites in metastatic renal cell carcinoma. *Future Oncol.* **2017**;13(22):1967–1980. doi:10.2217/fon-2017-0175
19. Mitsui Y, Shiina H, Kato T, et al. Versican promotes tumor progression, metastasis and predicts poor prognosis in renal carcinoma. *Mol Cancer Res.* **2017**;15(7):884–895. doi:10.1158/1541-7786.Mcr-16-0444
20. Walther MM, Alexander RB, Weiss GH, et al. Cytoreductive surgery prior to interleukin-2-based therapy in patients with metastatic renal cell carcinoma. *Urology.* **1993**;42(3):250–257. doi:10.1016/0090-4295(93)90612-e
21. de Leon AD, Pedrosa I. Imaging and screening of kidney cancer. *Radiol Clin North Am.* **2017**;55(6):1235–+. doi:10.1016/j.rcl.2017.06.007
22. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning users' guides to the medical literature. *JAMA J Am Med Assoc.* **2019**;322(18):1806–1816. doi:10.1001/jama.2019.16489
23. Singh NP, Bapi RS, K P. Vinod: machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. *Comput Biol Med.* **2018**;100:92–99. doi:10.1016/j.compbiomed.2018.06.030
24. Chen ST, Jiang LR, Gao F, et al. Zheng: machine learning-based pathomics signature could act as a novel prognostic marker for patients with clear cell renal cell carcinoma. *Br J Cancer.* **2022**;126(5):771–777. doi:10.1038/s41416-021-01640-2
25. Kim H, Lee SJ, Park SJ, Choi IY, Hong S-H. Hong: machine learning approach to predict the probability of recurrence of renal cell carcinoma after surgery: prediction model development study. *JMIR Med Inform.* **2021**;9(3):e25635. doi:10.2196/25635
26. Collins GS, Reitsma JB, Altman DG, M KG. Moons: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol.* **2015**;68(2):112–121. doi:10.1016/j.jclinepi.2014.11.010
27. Moons KGM, Altman DG, Reitsma JB, et al. Collins: transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and Elaboration. *Ann Internal Med.* **2015**;162(1):W1–W73. doi:10.7326/m14-0698
28. Capitanio U, Bensalah K, Bex A, et al. Epidemiology of Renal Cell Carcinoma. *Europ Urol.* **2019**;75(1):74–84. doi:10.1016/j.eururo.2018.08.036
29. Procházková K, Vodička J, Fichtl J. Outcomes for patients after resection of pulmonary metastases from clear cell renal cell carcinoma: 18 years of experience. *Urol Int.* **2019**;103(3):297–302. doi:10.1159/000502493
30. Margulis V, Shariat SF, Rapoport Y, et al. Wood: development of accurate models for individualized prediction of survival after cytoreductive nephrectomy for metastatic renal cell carcinoma. *Europ Urol.* **2013**;63(5):947–952. doi:10.1016/j.eururo.2012.11.040
31. Shinagare AB, Krajewski KM, Braschi-Amirfarzan M, H N. Ramaiya: advanced renal cell carcinoma: role of the radiologist in the era of precision medicine. *Radiology.* **2017**;284(2):332–350. doi:10.1148/radiol.2017160343
32. Wang W, Cao K, Jin S, Zhu X, Ding J, Peng W. Differentiation of renal cell carcinoma subtypes through MRI-based radiomics analysis. *Eur Radiol.* **2020**;30(10):5738–5747. doi:10.1007/s00330-020-06896-5
33. Selvi I, Demirci U, Bozdogan N, Basar H. The prognostic effect of immunoscore in patients with clear cell renal cell carcinoma: preliminary results. *Int. Urol Nephrol.* **2020**;52(1):21–34. doi:10.1007/s11255-019-02285-0
34. Wang JK, Zhanghuang CH, Tan XJ, et al. He: development and validation of a nomogram to predict distant metastasis in elderly patients with renal cell carcinoma. *Front Public Health.* **2022**;9:831940. doi:10.3389/fpubh.2021.831940
35. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med.* **2018**;46(4):547–553. doi:10.1097/ccm.0000000000002936
36. Zhang ZZ, Ho KM, C Y. Hong: machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Critical Care.* **2019**;23. doi:10.1186/s13054-019-2411-z
37. Chen C, Zhang QM, Ma Q, Yu B. LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom Intell Lab Syst.* **2019**;191:54–64. doi:10.1016/j.chemolab.2019.06.003
38. Wang DN, Li L, Zhao D. Corporate finance risk prediction based on LightGBM. *Inf Sci.* **2022**;602:259–268. doi:10.1016/j.ins.2022.04.058
39. Saito T, Rehmsmeier M, Brock G. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* **2015**;10(3):e0118432. doi:10.1371/journal.pone.0118432
40. Li Y, Chen P, Chen Z. A population-based study to predict distant metastasis in patients with renal cell carcinoma. *Ann Palliat Med.* **2021**;10(4):4273–+. doi:10.21037/apm-20-2481
41. Hutterer GC, Patard -J-J, Jeldres C, et al. Patients with distant metastases from renal cell carcinoma can be accurately identified: external validation of a new nomogram. *BJU Int.* **2008**;101(1):39–43. doi:10.1111/j.1464-410X.2007.07170.x