Classifying subtypes and predicting survival of renal cell carcinoma using histopathology image-based deep learning

Fei Wang
China Pharmaceutical University
Yuanzhe Geng
China Pharmaceutical University
Ting Wang
Nanjing University
Ke Zhao
The Affiliated Jiangyin Hospital of Southeast University Medical College
Bin Xu
Affiliated Zhongda Hospital of Southeast University
Dachuan Zhang
The Third Affiliated Hospital of Soochow University
Hongzhou Cai
The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research
Qinbo Yuan
Huaiyin People's Hospital of Huaian City
Hang Lu
China Pharmaceutical University
Yue Zhang
China Pharmaceutical University
Xu Li
China Pharmaceutical University
Yangyang Sun
China Pharmaceutical University
Hang Gong
China Pharmaceutical University
Raphael N Alolga
China Pharmaceutical University
Xiangshan Fan
Nanjing Drum Tower Hospital

Gaoxiang Ma
China Pharmaceutical University https://orcid.org/0000-0002-1774-6445
Lian-Wen Qi (Qilw@cpu.edu.cn)
China Pharmaceutical University https://orcid.org/0000-0003-0728-4475

Article

Keywords: Renal cell carcinoma, Deep learning, Histopathology, Diagnosis, Prognosis

Posted Date: May 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-533678/v1

Classifying subtypes and predicting survival of renal cell carcinoma using histopathology image-based deep learning

Fei Wang[1a], Yuanzhe Geng[1a], Ting Wang[2a], Ke Zhao[3a], Bin Xu[4a], Dachuan Zhang[5a], Hongzhou Cai[6a], Qinbo Yuan[7], Hang Lu[1], Yue Zhang[1], Xu Li[1], Yangyang Sun[1], Hang Gong[1], Raphael N. Alolga[1], Xiangshan Fan[2*], Gaoxiang Ma[8*], Lian-Wen Qi[1,8*].

From [1]Clinical Metabolomics Center, China Pharmaceutical University, Nanjing, 211198, China; [2]Department of Pathology, The Affiliated Drum Tower Hospital, Nanjing University Medical School, Nanjing, 210008, China; [3]Department of Pathology, The Affiliated Jiangyin Hospital of Southeast University Medical College, Wuxi, 214400, China; [4]Department of Urology, The Affiliated Zhongda Hospital of Southeast University, Nanjing, 210003, China; [5]Department of Pathology, The Third Affiliated Hospital of Soochow University, Changzhou, 213000, China; [6]Department of Urology, The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research, Nanjing, 210009, China; [7]Department of Urology, Huaiyin People's Hospital of Huai'an City, Huai'an, 223300, China; [8]State Key Laboratory of Natural Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing, 211198, China.

[a] These authors contributed to this work equally.

REPRINTS AND CORRESPONDENCE: Address reprint requests to Dr. Lian-Wen Qi at the Clinical Metabolomics Center of China Pharmaceutical University, E-mail address: Qilw@cpu.edu.cn; Dr. Gaoxiang Ma at the School of Chinese Herbal Medicines of China Pharmaceutical University, E-mail address: gaoxiang_ma@163.com; Dr. Xiangshan Fan at Department of Pathology, The Affiliated Drum Tower Hospital, Nanjing University Medical School, E-mail address: fanxiangshan@nju.edu.cn.

Keyword: Renal cell carcinoma; Deep learning; Histopathology; Diagnosis; Prognosis

Abstract

Classifying histopathological subtypes and predicting survival of renal cell carcinoma (RCC) patients are critical steps towards treatment. In this work, we first proposed a deep learning method involving patch-based segmentation, intelligent feature extraction and heatmap visualization for classifying RCC into clear cell RCC, papillary RCC, chromophobe RCC, and adjacent benign tissue. This algorithm was trained and validated using 2,374,446 patches, 6,340 whole-slide images, 2,399 patients from The Cancer Genome Atlas and 6 independent centers. The classifiers provided areas under the curves of 0.979 to 0.996 in the internal phase, and 0.914 to 0.995 in the 6-center external phase. Furthermore, a modified deep learning approach comprising automated detection of regions of interest, patch-level learning, and morphological features-based risk scoring was developed for predicting survival of clear cell RCC patients. The prognostication model provided a hazard ratio for poor versus good prognosis of 2.63 [95% confidence interval (CI) 1.53–4.50, P = 4.35e-4] in the testing set, and 2.57 [95% CI 1.43–4.64, P = 1.68e-3] in an independent validation set using multivariable analyses. In conclusion, the developed histopathology image-based deep learning frameworks have the clinical potential to assist pathologists in systematically evaluating histological information of RCC patients.

Introduction

Renal cell carcinoma (RCC) accounts for >90% of all kidney cancers. Epidemiological studies show that RCC represents approximately 2.2% of all cancers, with ~400,000 new cases and ~175,000 deaths yearly[1]. It can be classified into three major subtypes: clear cell RCC (ccRCC), the most common type accounting for 70% of all cases; papillary RCC (pRCC) which represents 15% to 20% of all cases, and chromophobe RCC (chRCC), that accounts for 5% of reported cases[2]. The remaining subtypes are very rare with each accounting for ⩽1% of total incidence. Each subtype of RCC has its specific histopathology, genetic characteristics, clinical course, and response to therapy[3].

Subtype classification and outcome prediction of RCC patients are critical steps towards precise treatment. Histopathological slide is the gold standard of RCC subtype and stage[2,4]. Classification and prognostication based on human assessment remain time-consuming and relatively subjective. In some cases, the distinction among RCC types is not clear as they may share non-specific morphological patterns[5]. In addition, RCC is an extremely heterogeneous disease, making prediction of prognosis a great challenge[6].

Deep learning has achieved impressive successes in digital image analysis[7]. It enables direct extraction of imaging features associated with classification and prognosis without explicit programming. Deep learning has been applied to histopathology-based segmentation, classification, and survival prediction of lung cancer[8], colorectal cancer[9], basal cell carcinoma[10], malignant mesothelioma[11], breast cancer[12], prostate cancer[13], glioma[14], and pan-cancer analysis[15]. Although few models have been used in clinical practice[16], deep learning algorithms hold considerable promise and enormous potential in disease diagnosis and prognosis.

Development of deep learning models to study renal cancers have been very few. Fenstermaker et al. trained a deep learning model to distinguish RCC subtypes based on biopsy images of 42 patients from The Cancer Genome Atlas (TCGA)[17]. Tabibu et al. proposed a deep learning method for the classification and survival prediction of RCC subtypes using images from TCGA[18]. Chen et al. developed a strategy for integrating histology image and genomic features to predict the outcomes of ccRCC patients from TCGA[19]. Marostica et al. diagnosed RCC histological subtypes and predicted stage I ccRCC patients' survival outcomes[20]. Evidently, these studies have a limitation in terms of small sample size and do not include multi-independent cohorts. Insufficient validation data hinder generalization of the underlying hypotheses, leading to a possibility of overfitting[21]. To overcome the complexity and heterogeneity of RCC, this work proposed histopathology imaged-based deep learning frameworks for type classification and clinical outcome prediction.

Results

Patient information.

The samples for classification covered 6,340 whole-slide images from 2,399 RCC patients, of which 3,260 slides of 941 patients were from TCGA and 3,080 slides of 1,458 patients were from 6 independent cohorts (Supplementary Table 1). Images from TCGA included 1,550 ccRCC, 678 pRCC, 304 chRCC, and 728 adjacent benign tissues. Images from the 6 independent cohorts included 1,872 ccRCC, 454 pRCC, 586 chRCC, and 168 adjacent benign tissues. Patient characteristics are shown in Supplementary Tables 2–8. Patients from TCGA were divided into 55% training set, 15% tuning set, and 30% internal testing set, and patients

from the 6 independent cohorts were used as external validation sets.

The samples for prognostication included 493 whole-slide images of 488 ccRCC patients from TCGA and 343 slides of 316 ccRCC patients from an independent TADTH cohort. The patient characteristics and follow-up information are shown in Supplementary Table 9. Inclusion criteria were age 18 years or older, ccRCC histologically proven to be R0 stage I–IV and grade 1–4. The median follow-up period of ccRCC patients was 39.0 months (0.1–151.2) for TCGA and 60.9 months (1.0–100.5) for TADTH. For TADTH cohort, follow-up consisted of telephone calls following the initial diagnosis. Of these, 72 (72/388, 18.6%) were excluded because of incomplete follow-up data. A total of 343 FFPE slides were obtained from the 316 patients (316/388, 81.4%; 87 women and 229 men). Clinicopathological findings were based on tumor–node–metastasis (TNM) classification. There were 219 patients with stage I ccRCC, 31 with stage II, 59 with stage III, and 7 patients with distant metastases classified as stage IV ccRCC.

HistoQC analysis was employed to check the image quality in terms of a number of metrics and features (Supplementary Fig. 2a). These include microns per pixel (Supplementary Fig. 2b), brightness (Supplementary Fig. 2c) and contrast (Supplementary Fig. 2d), etc. Of the 6,384 whole-slide images from TCGA and the independent testing set, 44 slides were excluded for being unsuitable for subsequent computational analysis.

Development of classifier models.

The construction of classifier models involved steps of patch-based segmentation, intelligent feature extraction, and heatmap visualization (Fig. 1). Slides and patches information for each set are shown in Supplementary Table 10. Inception V3 achieved a bit better performance than

the other four state-of-the-art CNN architectures, including VGG19, DenseNet, ResNet, and MobileNet (Supplementary Table 11). Models trained at 5× provided higher performance compared with 10× and 20× (Supplementary Table 12). Models trained with Inception V3 at 5× were then employed for subsequent analysis.

In the internal testing set, a binary classier achieved an AUC of 0.996 [0.990–1.000] at the patient level using FFPE slides, and 0.993 [0.984–0.998] for frozen slides (Supplementary Table 12) in differentiating tumor from normal tissues. In the classification of RCC subtypes, a three-way classification provided a macro-average AUC of 0.988 [0.979–0.995], a micro-average AUC of 0.987 [0.979–0.995] for the FFPE samples, and a macro-average AUC of 0.979 [0.960–0.996], a micro-average AUC of 0.983 [0.970–0.995] for the frozen samples. The performances of classifier models were also satisfactorily evaluated by Youden index (Supplementary Table 13), PRC (Supplementary Fig. 3a–d), confusion matrices (Supplementary Fig. 4a–d), and PDI (Supplementary Table 14).

Validation of the classifier models in 6 independent cohorts.

The classifier models were validated with 3,080 slides from 1,458 patients in the 6 multi-center external phase. In identifying malignancy from adjacent benign tissues, the binary classifier provided an AUC of 0.995 [0.987–0.999] at the patient level for TADTH cohort (Fig. 2a). In differentiating between the 3 RCC types, the trained three-way classifier generated a macro-average AUC of 0.968 [0.957–0.978] and a micro-average AUC of 0.967 [0.957–0.975] for FFPE slides from TADTH (Fig. 2b). For frozen slides from TADTH, it yielded a macro-average AUC of 0.916 [0.817–0.989] and a micro-average AUC of 0.914 [0.835–0.977] (Fig. 2c). The three-way classifier offered a macro-average AUC of 0.964 [0.937–0.987] and a micro-average

AUC of 0.947 [0.919–0.971] for FFPE slides from TAJH (Fig. 2d), 0.960 [0.905–0.994] and 0.965 [0.934–0.989] from TAZH (Fig. 2e), 0.958 [0.914–0.986] and 0.951 [0.912–0.984] for FFPE slides from TTAH (Fig. 2f), 0.964 [0.907–0.999] and 0.975 [0.945–0.995] for TACH (Fig. 2g), and 0.958 [0.900–0.991] and 0.937 [0.872–0.980] for HPH (Fig. 2h). The performances of classifier models were also satisfactorily evaluated by Youden index (Supplementary Table 13), PRC (Supplementary Fig. 3e–l), confusion matrices (Supplementary Fig. 4e–l), and PDI (Supplementary Table 14).

Heatmaps of the prediction probability over slides were generated to discern the tumor regions associated with histological patterns. We have shown representations of images of the four types (Fig. 3a). Heatmap visualizations were produced for which color is proportional to the predicted probability (0~1) of the patch. The generated heatmap clearly differentiated tumor from normal region at the patch level (Fig. 3b) by a binary classier, and then classified the three RCC subtypes of tumor patches by a three-way classifier (Fig. 3c). Most of the predicted patches had a strong true positive probability for a certain type of RCC (Fig. 3d).

Comparison of the classifier with pathologists.

A subset of 221 FFPE and frozen slides from the internal testing set was used for algorithm-pathologist comparison[22]. These included 24 out of 32 misclassified cases and 197 out of 492 correctly classified by the classifier. Pathologist 1 with 3 years' experience and pathologist 2 with 5 years' experience from TADTH were instructed to assess the digital whole-slide images alone, independently of the clinical data provided by TCGA. The pathologists were free to use a slide analytical software (ASAP; version 1.8) at varying zoom levels of up to 40× magnification. For FFPE slides, the agreement of the classifier with TCGA using Cohen's

Kappa statistic was comparable with the pathologists (0.839 versus 0.758 for pathologist 1 and 0.813 for pathologist 2) (Supplementary Fig. 5, Supplementary Table 15). For frozen slides, the performance of the classifier was also comparable with the pathologists (0.820 versus 0.709 for pathologist 1 and 0.767 for pathologist 2) (Supplementary Fig. 5, Supplementary Table 15).

Of the images misclassified by the classifier, 83% (20/24) were also incorrectly diagnosed by at least one pathologist, while 60% (30/50) of those misclassified by at least one pathologist were classified successfully by the classifier. A few slides showed mixed histologic features, poor differentiation or massive necrosis (Supplementary Fig. 6), leading to possible misclassification by the model and the two pathologists.

Development of the prognostic model.

The ccRCC is the most common subtype and accounts for the majority of deaths. To obtain a prognostic biomarker for ccRCC patients, a modified deep learning architecture using CNN combined with Cox regression was developed. Patients in the testing set were divided into poor and good prognostic groups based on the median risk score obtained from the training set. The prognostication framework consisted of automated detection of regions of interest, patch-level learning and morphological features-based risk scoring (Fig. 4). VGG19 generated a bit better performance than Inception V3, DenseNet, ResNet, and MobileNet (Supplementary Table 16). We compared the performances of 10, 20 and 30 representative image patches with the largest average nuclei size for each slide. We observed that the c-index value did not increase with more patches but slightly deceased using 20 and 30 patches compared to 10 patches (Supplementary Table 17). As a result, we used 10 image patches of each slide as inputs to VGG19 for subsequent prognostication.

Our survival deep-learning model showed strong prognostic power with a c-index of 0.779, outperforming manual histologic-grade system of 0.748. Kaplan–Meier analysis showed that the prognostic model had better prediction capability than histologic grade (log rank $P = 3.49e\text{-}5$ versus $P = 2.22e\text{-}3$) (Fig. 5a,b). The risk index (low risk, high risk) was a prognostic factor for overall survival in a univariate Cox analysis (HR 2.93 [95% CI 1.72–5.00], $P = 7.84e\text{-}5$) (Supplementary Table 18). In multivariate analysis, the risk index was independently prognostic of survival in all-stage (HR = 2.63 [95% CI 1.53–4.50], $P = 4.35e\text{-}4$), early-stage (HR = 2.41 [95% CI 1.03–5.61], $P = 0.04$), and late-stage (HR = 2.81 [95% CI 1.38–5.72], $P = 4.30e\text{-}3$) tumors adjusting for stage and age (Fig. 5c). The histologic grade (G1+G2, G3+G4) was only prognostic in all-stage but not in early-stage and late-stage tumors. The predicted risk score was highly consistent with patient outcomes, especially for 3-year follow-up (Fig. 5d). The model's performance remained satisfactory as measured by accelerated failure time analyses (Supplementary Table 19). The predicted risk scores were significantly associated with tumor grading and staging (Spearman's $\rho = 0.34$ and 0.22, respectively, both $P < 0.01$) (Supplementary Table 20). The established scoring algorithm allowed patches to be scored for each patient. Comparison between high-risk patches and low-risk patches was beneficial in obtaining morphological characteristics associated with prognosis (Supplementary Fig. 7).

Validation of the prognostic model in an independent cohort.

We validated the model in an independent cohort of 316 RCC patients from TADTH. The model showed a c-index of 0.769, better than manual histologic grade of 0.753. This prognostic model provides significant survival differences between high- and low-risk groups (log rank $P = 5.88e\text{-}4$), better than the histologic grade (log rank $P = 1.02e\text{-}3$) (Fig. 5e,f). The risk index was a

predictor of patient outcomes in a univariate Cox analysis (HR 2.70 [95% CI 1.50–4.87], P = 9.65e-4) (Supplementary Table 18). In multivariate analysis, the risk index was independently prognostic of survival in all-stage (HR = 2.57 [95% CI 1.43–4.64], P = 1.68e-3), early-stage (HR = 2.30 [95% CI 1.07–4.94], P = 0.03), and late-stage tumors (HR = 2.65 [95% CI 1.03–6.83], P = 0.04) (Fig. 5g). The histologic grade showed prognostication in all-stage but not in early-stage and late-stage tumors. We have presented the distribution plot of risk scores and survival time of ccRCC patients in Fig. 5h. The performance in independent cohort was also evaluated by accelerated failure time analyses (Supplementary Table 19). The risk index was also positively correlated with tumor grade and stage (Spearman's ρ = 0.45 and 0.16, respectively, $P < 0.01$). (Supplementary Table 20).

Discussion

In this study, we developed histopathology image-based deep learning frameworks: a classifier model with Inception V3 to classify RCC subtypes, and a prognostication model with VGG19 to predict outcomes of ccRCC. The classifier model involves patch-based segmentation, intelligent feature extraction, and heatmap visualization using whole-slide images. It was trained and validated using 3,260 slides of 941 patients from TCGA and 3,080 real-world slides of 1,458 patients from 6 independent centers. The algorithm could unambiguously distinguish normal from tumor tissues with AUC >0.99 and effectively distinguished between the three RCC subtypes with AUC >0.91, with accuracy comparable to or even better than two clinical pathologists. To obtain a prognostic biomarker, the developed classifier model first automatically outlined cancerous regions of ccRCC, and then the prognostication model was

219 built by combining deep learning with Cox regression. Independent validation confirmed that
220 the predicted risk index is significantly associated with clinical outcome of ccRCC.

221 In general, most cases of RCC can be easily classified based on histological criteria. We
222 observed that the misclassified cases displayed a combination of morphological features.
223 Indeed, the presence of clear cells is not unique to ccRCC but can also be observed in some
224 cases of pRCC and chRCC. Similarly, papillary structures characteristic of pRCC, can also be
225 present in other subtypes[23].

226 When applied to external validation, there was a slight drop of the classifier' performance
227 in differentiating tumor from normal (AUC 0.1%) and subtypes (AUC 1%–7%). It should be
228 noted that these slides were collected without data curation. The interlaboratory differences in
229 slide preparation might explain the decrease. The difference between the scanners in brightness
230 and contrast could also somewhat affect the prediction accuracy.

231 ccRCC is the most common variant and has wide ranging clinical outcomes attributed to its
232 tumor heterogeneity[6]. Historically, the most widely used histological grading system for ccRCC
233 is the Fuhrman grading system[23], which relies on the experience of pathologists. In recent years,
234 automatic models depending on extracting engineered imaging features have been proposed[18,24].
235 More recently, survival analysis has been approached with a classification task by predicting
236 several periods of survival time divided by specific time points[20]. These classifiers, however,
237 cannot model the risk values with certain survival times and lack independent follow-up cohort.
238 This work predicted the patient outcomes by learning prognostic features directly from slides.
239 It employed Cox regression, a time-to-event model that can better fit the prediction of survival
240 and model all patients' risks for a range of survival time.

Our prognostic model highlights the importance of dedifferentiation, stroma component, cellular diversity, inflammation, and growth pattern for prognosis. We observed that a considerable proportion of high-risk patches were located in the sarcomatoid and rhabdoid regions, which are well known as malignant and dedifferentiated components[24,26]. Some contained pleomorphic or giant cells and coagulative necrosis, consistent with previous observations[27]. Some had unique characteristics of infiltrative growth patterns[28]. Others presented stromal components in the tumor microenvironment[29]. In contrast, good prognostic patches were located in compact small/large nests and cystic regions. Some showed intratumoral inflammatory reaction, tubular, clear cell papillary or chromophobe cell-like patterns[30].

This study has some limitations. The algorithms focused on classifying the three common subtypes of renal cancer, and did not identify rare subtypes such as Mit family translocation RCC and clear cell papillary RCC due to unavailability of samples. For clinical application of the model, more samples are encouraged to fine-tune the algorithm to increase its generalizability. Another limitation is that our diagnostic algorithms were only trained on H&E-stained images. Integrating features from immunochemical-stained images remains to be explored.

In conclusion, we have presented deep learning frameworks for systematic assessment of histological information on RCC. The established models have been evaluated in multi-center, independent cohorts. Our models have the potential of accelerating the pathologist's evaluation for kidney tissue slides, and making it more objective and replicable in clinic.

Methods

Data sources.

Our biospecimen were collected from TCGA Resource Network3[1] and 6 independent cohorts in China. Based on a previous research[32], 15 initial samples originally submitted as ccRCC in TCGA were reclassified as chRCC. The 6 independent cohorts were: The Affiliated Drum Tower Hospital (TADTH), The Affiliated Jiangyin Hospital of Southeast University Medical College (TAJH), The Affiliated Zhongda Hospital of Southeast University (TAZH), The Third Affiliated Hospital of Soochow University (TTAH), The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research (TACH), and Huaiyin People's Hospital of Huai'an City (HPH). Both formalin-fixed and paraffin-embedded (FFPE) sections and frozen sections were included. There were no exclusion criteria on age, gender or race.

Cases were reviewed by two pathologists with more than 3 years' experience. The diagnoses were made based on morphology following the World Health Organization recommendation. All slides were finally checked by Prof. Fan Xiangshan, a pathologist with many years of experience and the head of pathology department at TADTH. In case of inconsistent decisions, immunostaining was further performed for classification using a panel of antibodies.

Slides from TTAH were scanned by a KF-PRO-120 scanner (Konfoong Biotechnology) at 20× magnification, and samples from the other independent cohorts were scanned by a NanoZoomer 2.0-RS scanner (Hamamatsu). Quality checking was performed for the whole-slide images by an open-source tool *HistoQC* to exclude unsuitable samples[33].

Construction of classification models.

Several commonly used convolutional neural networks (CNNs) were compared including Inception V3, 19-layer Visual Geometry Group (VGG19), DenseNet, ResNet, and MobileNet. We optimized the parameters of fully connected layers as well as the convolution layers by unfreezing the whole network. The parameters of each network were initialized with pretrained parameters from the ImageNet classification challenge as a starting point. For every image in the training set, rotational invariance was achieved through data augmentation with random horizontal and vertical flips before being fed into the network. For the training process, the batch size was set to 32. The cross-entropy loss was set as the loss function on the training and tuning sets. All parameters of the networks were trained jointly using stochastic gradient descent[34] as a backpropagation method with learning rate of 0.05, weight decay of 0.9 and momentum of 0.9. The number of training epochs was 50, and the network with the lowest loss on the tuning set was selected.

Patients of each RCC type from TCGA dataset were divided into 55% training set, 15% tuning set, and 30% internal testing set. For the training and tuning sets, tumor lesions in each malignant slide were manually annotated by pathologists (Supplementary Fig. 1). Each whole-slide image was automatically split into hundreds of nonoverlapping 299×299-pixel patches using the Openslide (version 3.4.1). Benign patches for training and tuning sets were sampled from both adjacent benign tissue slides and non-tumor regions in malignant slides. Segmentation at magnifications of 5×, 10× and 20× were compared. Patches from patients in the training set were input into the architectures to intelligently extract morphological features and make decisions with parameters constantly optimized. The tuning set was then employed

to evaluate the trained algorithm and avoid overfitting. After tuning, the algorithm was then kept locked in subsequent validations.

For internal and external validation, the slides were automatically segmented into patches without manual annotation. Each patch was predicted as tumor or benign by a binary classifier with a possibility threshold of 0.5. Tumor patches were then classified as ccRCC, pRCC or chRCC based on the possibility predicted by a three-way subtype classifier. Predictions were made with probability and indicated using colors, producing a heatmap for visualization of each slide. The prediction results were then aggregated by calculating percentage of normal/tumor patches over all patches, or percentage of one-RCC type patches over predicted tumor patches in slides from each patient.

Establishment of a prognostic model.

ccRCC patients from TCGA were randomly divided into 55% training set, 15% tuning set and 30% internal testing set. An independent cohort from TADTH was used for external validation. Slides were segmented into nonoverlapping 1196×1196-pixel patches at 20×. The developed classifier model first automatically outlined cancerous patches from benign ones in each ccRCC slide. The nuclei size of cancerous patches was calculated by a Python package (https://github.com/DigitalSlideArchive/HistomicsTK). Top 10, 20 or 30 tumor patches with the largest average nuclei size for each slide were selected and compared.

Subpatches (224×224-pixel) were further randomly sampled from the representative patches as inputs. The prognostic deep learning model was trained end-to-end, directly from slide to survival time. We adapted the predictive algorithm by combining CNN architecture and Cox regression model. The performances of CNNs were compared, including VGG19,

Inception V3, DenseNet, ResNet, and MobileNet. The CNN made use of various convolutions from original architecture to extract image features. The last layer was changed to a fully connected layer containing one node, which predicted a risk score for the input subpatch. The risk scores in each subpatch were presented to a Cox proportional hazards layer allowing for the use of censored data to calculate the Cox loss function. The loss function was optimized using stochastic gradient descent with a learning rate of 5.0e-5, weight decay of 0.9 and momentum of 0.9. The training process was run for 100 epochs, and the model with the lowest loss on the tuning set was saved.

For internal and external validation sets, each representative tumor patch was split into nonoverlapping 16 subpatches. The median risk of these subpatches was calculated as the value for each patch, and then the highest value among all representative patches was selected as a patient risk. These processing procedures were designed to address tumor heterogeneity by emulating histological assessment by pathologists. In routine clinical practice, the pathologically assigned grade is based on the most malignant region when ccRCC shows morphologic grade variation[35]. The patients were then separated into equivalent groups of low- and high-risk on the basis of their predicted risk scores with the median risk in the training set as the cut-off point.

Statistics.

To evaluate the performance of classifiers, areas under the receiver operating characteristic (ROC) and precision-recall curves (PRC), Youden index, and polytomous discrimination index (PDI)[36] were computed using the Python library sklearn. Confidence intervals (CIs) 95% were measured using 1,000 iterations of the bootstrap method. The agreements between the classifier

and pathologists were measured by the Cohen's Kappa statistic[37]. For survival prediction, overall survival time was defined as the time from nephrectomy to death by any cause or the date of last follow-up. Harrell's c-index was used to quantify the concordance between predicted risks and true survival time. The survival curves of the predicted poor and good survival subgroups were plotted by the Kaplan-Meier method. The Cox regression model was used to obtain hazard ratios (HRs) and 95% CIs. Associations between the models and other risk factors were evaluated by Spearman's correlation coefficients. A two-sided *P* value of less than 0.05 was considered significant. All statistical analyses were done in R unless otherwise noted (R version 3.6.1). The open-source deep learning framework Keras (https://keras.io) was used to train and evaluate the algorithms.

Data availability

The publicly shared RCC histopathology images in TCGA dataset to train and test the models is available at https://portal.gdc.cancer.gov/. The dataset consists of 3,260 whole-slide images from 941 patients, including 1,550 ccRCC, 678 pRCC, 304 chRCC, and 728 adjacent benign tissues. The independent datasets are not publicly available due to hospital regulations and patient privacy. Source data are provided with this paper. The remaining data are available within the Article, Supplementary information, or available from the authors upon request.

Code availability

The code is available at https://github.com/cilcmc/dlrcc.

Author contributions

L.-W.Q., G.M., and X.F. conceived and directed the project. F.W. and Y.G., implemented and trained the deep learning model. H.L., Y.Z., X.L., Y.S., and H.G. collected and scanned slides from the independent datasets. F.W. and G.M. contributed to the analysis of the data. T.W., K.Z., B.X., D.Z., H.C., and Q.Y. helped identify and label the slides. Y.Z. and H.G. performed the follow up and collected the prognostic information of patients. F.W., L.-W.Q., G.M. and R.N.A. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Declaration of interests

The authors declare no competing interests.

Source of funding

This work was supported by the National Natural Science Fund of China for Distinguished Young Scholars (No. 81825023).

References

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. Moch, H., Cubilla, A.L., Humphrey, P.A., Reuter, V.E. & Ulbright, T.M. The 2016 WHO classification of tumours of the urinary system and male genital organs—part A: renal, penile, and testicular tumours. *Eur. Urol.* **70**, 93–105 (2016).
3. Escudier, B. et al. Renal cell carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **27**, v58–v68 (2016).

4. Leibovich, B.C. et al. Predicting oncologic outcomes in renal cell carcinoma after surgery. *Eur. Urol.* **73**, 772–780 (2018).
5. Hsieh, J.J. et al. Renal cell carcinoma. *Nat. Rev. Dis. Primers* **3**, 17009 (2017).
6. Gulati, S. et al. Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. *Eur. Urol.* **66**, 936–948 (2014).
7. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
8. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
9. Skrede, O.-J. et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350–360 (2020).
10. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
11. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
12. Bejnordi, B.E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
13. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
14. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA.* **115**, E2970–E2979 (2018).
15. Kalra, S. et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ. Digit. Med.* **3**, 31 (2020).
16. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
17. Fenstermaker, M., Tomlins, S.A., Singh, K., Wiens, J. & Morgan, T.M. Development and validation of a deep-learning model to assist with renal cell carcinoma histopathologic interpretation. *Urology* **144**, 152–157 (2020).

18. Tabibu, S., Vinod, P. & Jawahar, C. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Rep.* **9**, 1–9 (2019).
19. Chen, R.J. et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* (2020).
20. Marostica, E. et al. Development of a histopathology informatics pipeline for classification and prediction of clinical outcomes in subtypes of renal cell carcinoma. *Clin. Cancer Res.* (2021).
21. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
22. van Smeden, M., Van Calster, B. & Groenwold, R.H.H. Machine learning compared with pathologist assessment. *JAMA* **319**, 1725–1726 (2018).
23. Fuhrman, S.A., Lasky, L.C. & Limas, C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am. J. Surg. Pathol.* **6**, 655–664 (1982).
24. Cheng, J. et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res.* **77**, e91–e100 (2017).
25. Cheville, J.C. et al. Sarcomatoid renal cell carcinoma: an examination of underlying histologic subtype and an analysis of associations with patient outcome. *Am. J. Surg. Pathol.* **28**, 435–441 (2004).
26. Singh, R.R. et al. Intratumoral morphologic and molecular heterogeneity of rhabdoid renal cell carcinoma: challenges for personalized therapy. *Mod. Pathol*. **28**, 1225–1235 (2015).
27. Delahunt, B. et al. Grading of clear cell renal cell carcinoma should be based on nucleolar prominence. *Am. J. Surg. Pathol.* **35**, 1134–1139 (2011).
28. Cai, Q. et al. Ontological analyses reveal clinically-significant clear cell renal cell carcinoma subtypes with convergent evolutionary trajectories into an aggressive type. *EBioMedicine* **51**, 102526 (2020).
29. Junttila, M.R. & de Sauvage, F.J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501**, 346–354 (2013).
30. Verine, J. et al. Architectural patterns are a relevant morphologic grading system for clear cell renal cell carcinoma prognosis assessment: Comparisons with WHO/ISUP grade and integrated staging systems. *Am. J. Surg. Pathol.* **42**, 423–441 (2018).

31. Grossman, R.L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
32. Ricketts, C.J. et al. The Cancer Genome Atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* **23**, 313–326.e315 (2018).
33. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inform* **3**, 1–7 (2019).
34. Bottou, L. Stochastic gradient descent tricks. in *Neural networks: Tricks of the trade* 421–436 (Springer, 2012).
35. Delahunt, B. et al. A novel grading system for clear cell renal cell carcinoma incorporating tumor necrosis. *Am. J. Surg. Pathol.* **37**, 311–322 (2013).
36. Van Calster, B. et al. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat. Med.* **31**, 2610–2626 (2012).
37. McHugh, M.L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).

Figures

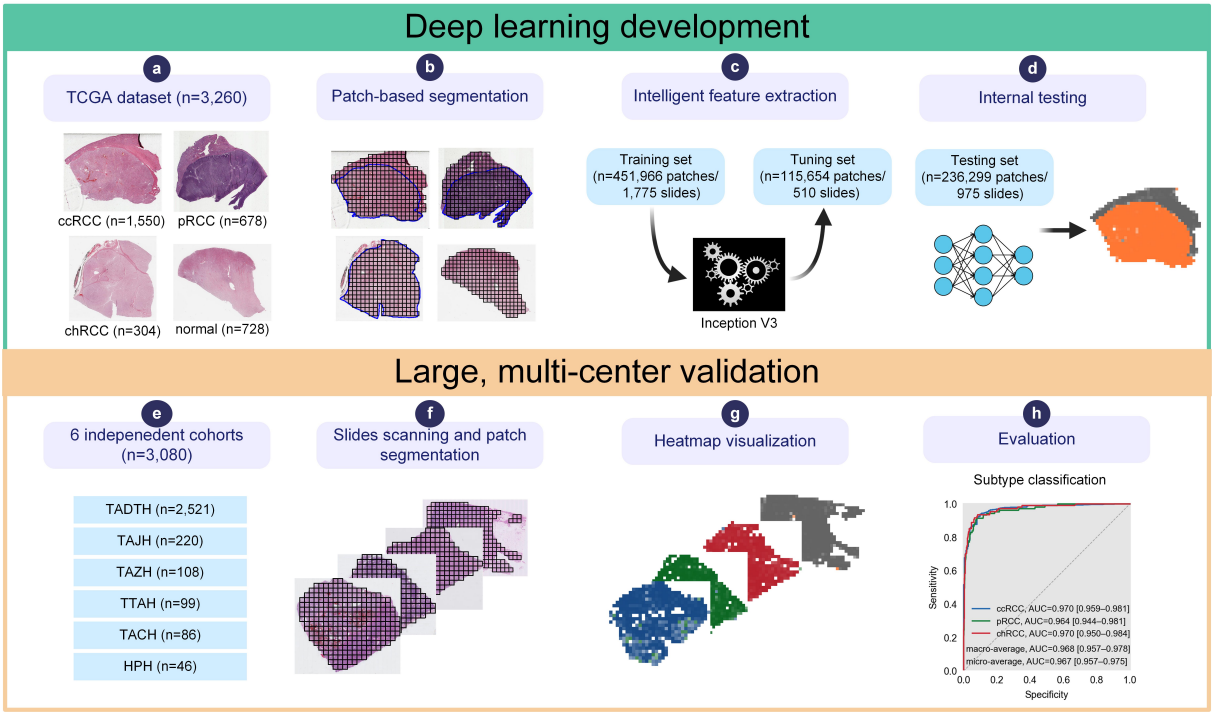


Fig. 1. Schematic presentation of the diagnostic system. a–d, Deep learning development. **a**, Whole-slide images of renal cell carcinoma (RCC) from The Cancer Genome Atlas (TCGA). A total of 3,260 images were collected, including 1,550 clear cell RCC (ccRCC), 678 papillary RCC (pRCC), 304 chromophobe RCC (chRCC), and 728 normal tissues. **b**, Manual annotation of tumor lesions for each slide and automatic segmention by nonoverlapping 299 × 299-pixel patches at 5× magnification with background removed. **c**, Training Inception V3 architecture to intelligently extract diagnostic features. The training set was input into the algorithm to make prediction, and the tuning set was used to avoid overfitting. **d**, Patch-based classification for the internal testing set. The results were aggregated per slide to generate the heatmaps. **e–h**, Large, multi-center validation. **e**, Assembling real-world slides from 6 independent cohorts. The framework was validated with 3,080 slides from The Affiliated Drum Tower Hospital (TADTH), The Affiliated Jiangyin Hospital of Southeast University Medical College (TAJH), The Affiliated Zhongda Hospital of Southeast University (TAZH), The Third Affiliated Hospital of Soochow University (TTAH), The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research (TACH), Huaiyin People's Hospital of Huai'an City (HPH). **f**, Slides scanning and patch segmentation. The slides were scanned and automatically segmented without manual annotation. **g**, Heatmap visualization. Patch-level diagnosis were made with predicted probability and are indicated using colors, producing a heatmap for visualization of each slide. **h**, Receiver operating characteristic and precision-recall curves, Youden index, and polytomous discrimination index statistics for classifications of normal tissue versus tumor and between RCC subtypes.

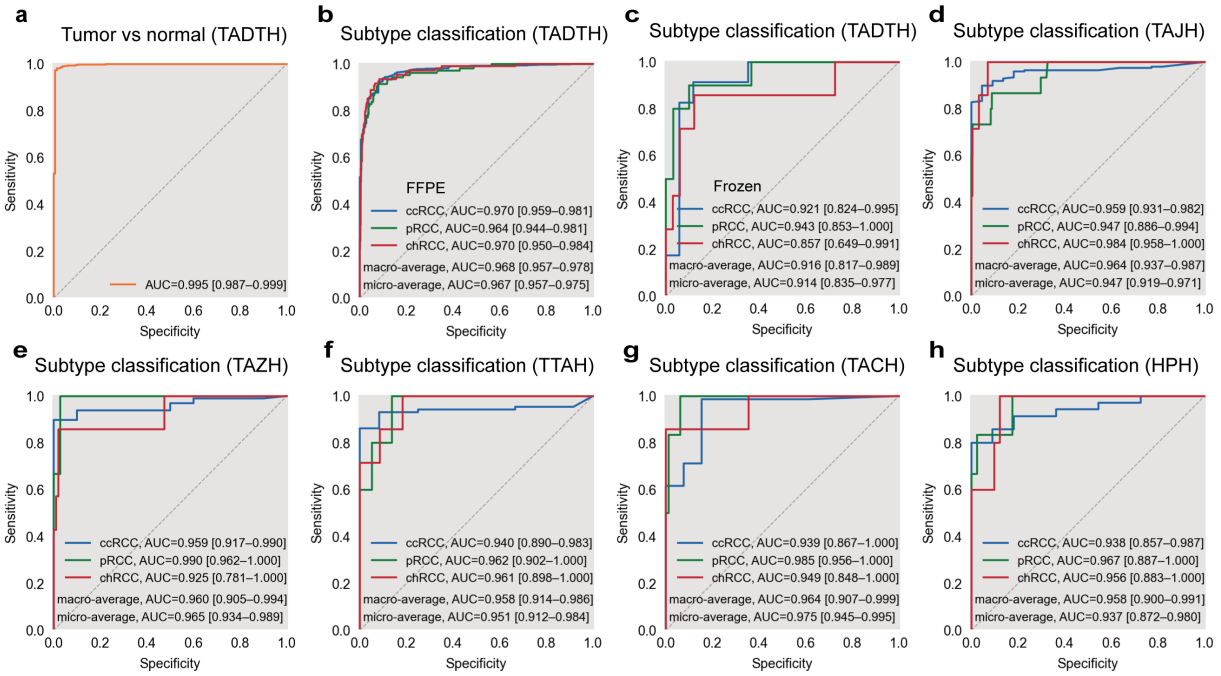


Fig. 2. Differential diagnosis of normal tissue from tumor and between renal cell carcinoma (RCC) types at the patient level. a. Receiver operating characteristic curves (ROC) for identifying tumor versus normal using FFPE slides from The Affiliated Drum Tower Hospital (TADTH, n=2,474 slides from 859 patients). **b**, ROC for classifying RCC types using FFPE slides from TADTH (n=2,306 slides from 859 patients). **c**, ROC for classifying RCC subtypes using frozen slides from TADTH (n=47 slides from 40 patients). **d**, ROC for classifying RCC types using FFPE slides from The Affiliated Jiangyin Hospital of Southeast University Medical College (TAJH, n=220 slides from 220 patients). **e,** ROC for classifying RCC subtypes using FFPE slides from The Affiliated Zhongda Hospital of Southeast University (TAZH, n=108 slides from 108 patients). **f**, ROC for classifying RCC subtypes using FFPE slides from The Third Affiliated Hospital of Soochow University (TTAH, n=99 slides from 99 patients). **g,** ROC for classifying RCC subtypes using FFPE slides from The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research (TACH, n=86 slides from 86 patients). **h**, ROC for classifying RCC subtypes using FFPE slides from Huaiyin People's Hospital of Huai'an City (HPH, n=46 slides from 46 patients).

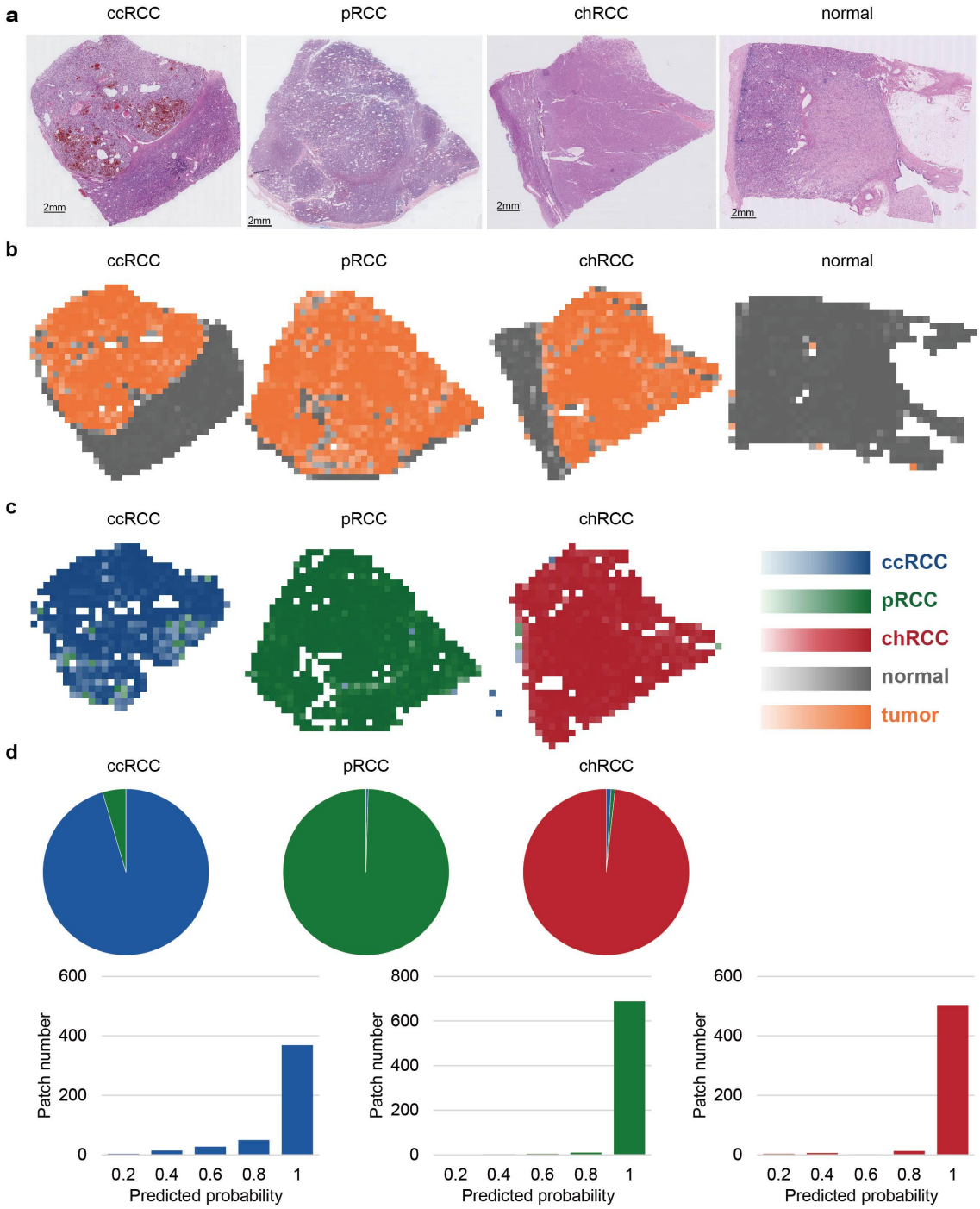
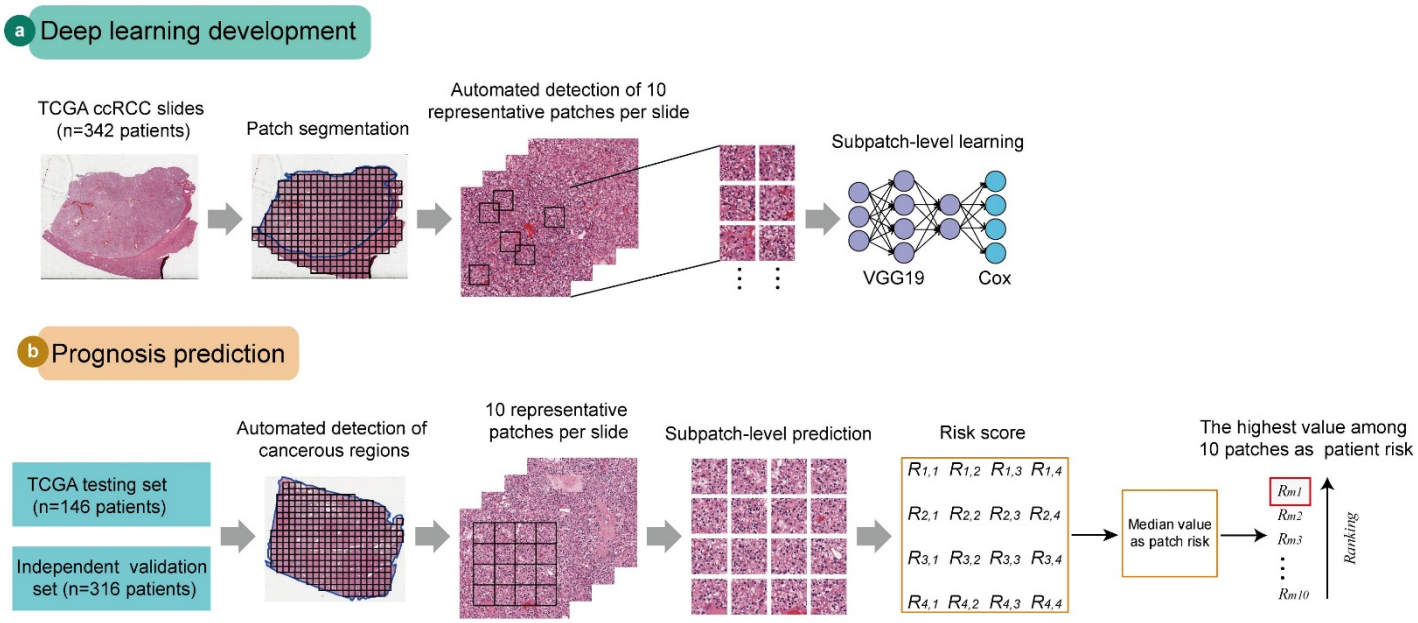


Fig. 3. Heatmap visualization of renal cell carcinoma (RCC) slides. a, Representative whole-slide images of the three types and adjacent benign tissues from The Affiliated Drum Tower Hospital (TADTH) cohort. **b**, Binary classifier for heatmap visualization of tumor versus normal. **c**, Three-way classifier for heatmap visualization of ccRCC versus pRCC versus chRCC. **d**, Patch-based prediction probability for a slide. The pie chart shows the predicted patch number for each category, and the bar chart shows the predicted probability of each patch. In **b**–**d**, the predicted heatmaps with probabilities are assigned to each patch, where grey is for patches classified as normal, orange for tumor, blue for ccRCC, green for pRCC, and red for chRCC.



5_ .

525 **Fig. 4. Pipeline of the prognostic framework. a**, Deep learning development. Whole-slide
526 images of clear cell renal cell carcinoma (ccRCC) patients from The Cancer Genome Atlas
527 (TCGA) dataset were randomly divided into training (55%), tuning (15%) and testing (30%)
528 sets. After patch-based segmentation, the classifier model first automatically outlined cancerous
529 regions of ccRCC. Ten representative patches (1196×1196-pixel at 20× magnification) with the
530 largest average nuclei in tumor regions were selected from each slide. Subpatches (224×224-
531 pixel) were sampled from the representative patches. The 19-layer Visual Geometry Group
532 (VGG19) architecture combined with Cox regression was employed to predict patient outcome
533 end-to-end. **b**, Prognosis prediction. For slides in the TCGA testing and independent validation
534 sets, cancerous regions were automatically detected by the diagnostic model, and 10
535 representative patches were selected from each slide. Each representative patch was split into
536 16 nonoverlapping subpatches; the median subpatch risk was calculated for each patch, and
537 then the highest value among 10 patches was selected as patient risk.
538

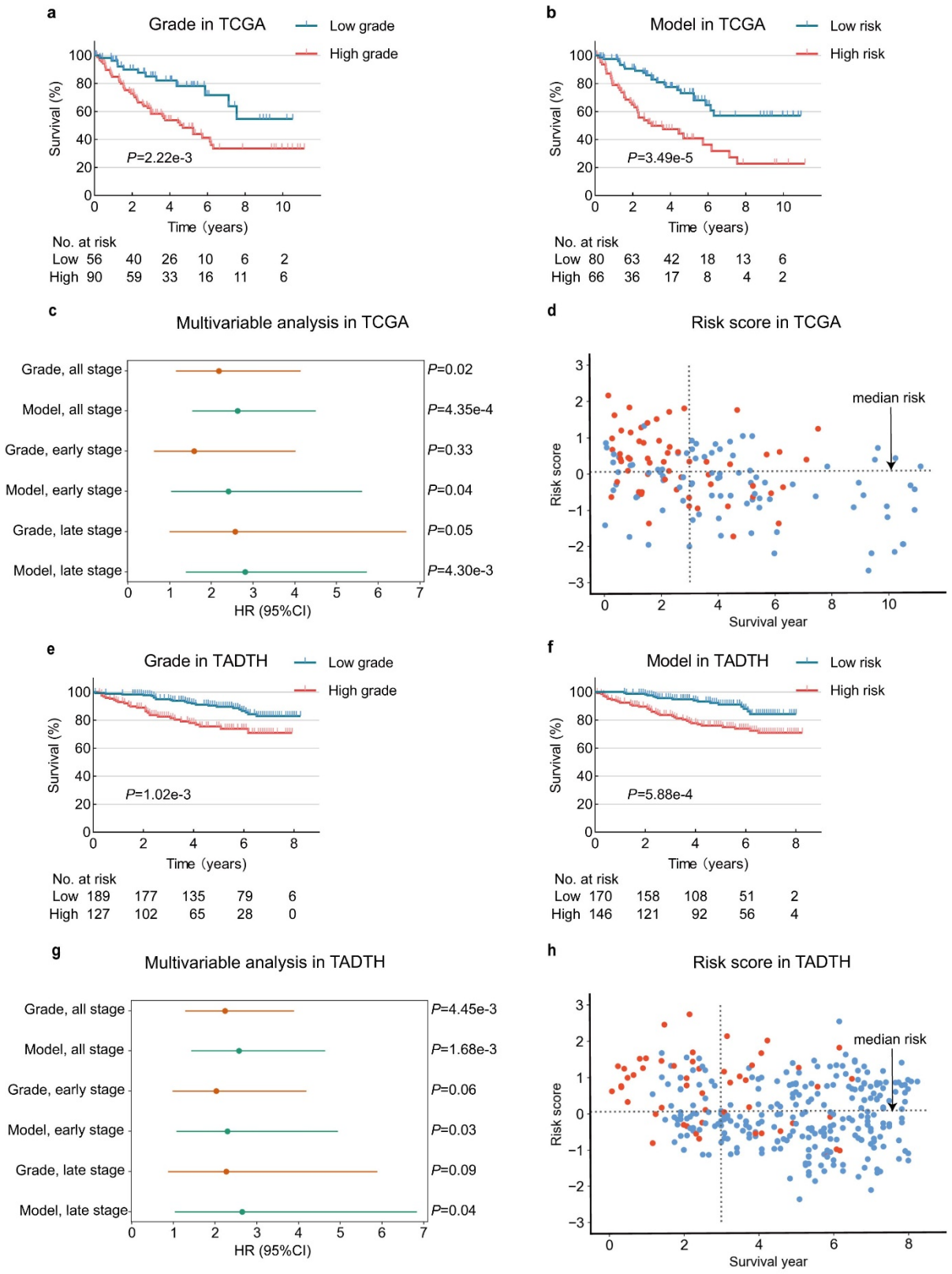


Fig. 5. Performances of prognostic model and manual histologic grade for clear cell renal cell carcinoma (ccRCC) patients. a, Kaplan-Meier analysis of histologic grade in The Cancer Genome Atlas (TCGA) testing set (n=146 patients). **b**, Kaplan-Meier analysis of the prognostic model in TCGA testing set. **c**, Hazard ratios (HRs) with 95% confidence interval (CI) of grade (G1+G2, G3+G4) and prognostic model (low risk, high risk) in multivariable analyses adjusting

for stage and age across different stages in TCGA testing set. **d**, The distribution of risk scores and outcomes of ccRCC patients in TCGA testing set. **e**, Kaplan-Meier analysis of histologic grade in The Affiliated Drum Tower Hospital (TADTH) independent validation set (n=316 patients). **f**, Kaplan-Meier analysis of the prognostic model in TADTH. **g**, HRs with 95% CI of histologic grade (G1+G2, G3+G4) and prognostic model (low risk, high risk) in multivariable analyses adjusting for stage and age across different stages in TADTH. **h**, The distribution of risk scores and outcomes of ccRCC patients in TADTH. The blue dots represent censored data and the red dots represent the dead patients. Prognostic model was highly associated with patient survival time, especially for a 3-year follow-up (**d**, **h)**. The median risk score obtained in the TCGA training set was used as threshold to divide patients into low- and high-risk groups (**a**–**h)**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.pdf