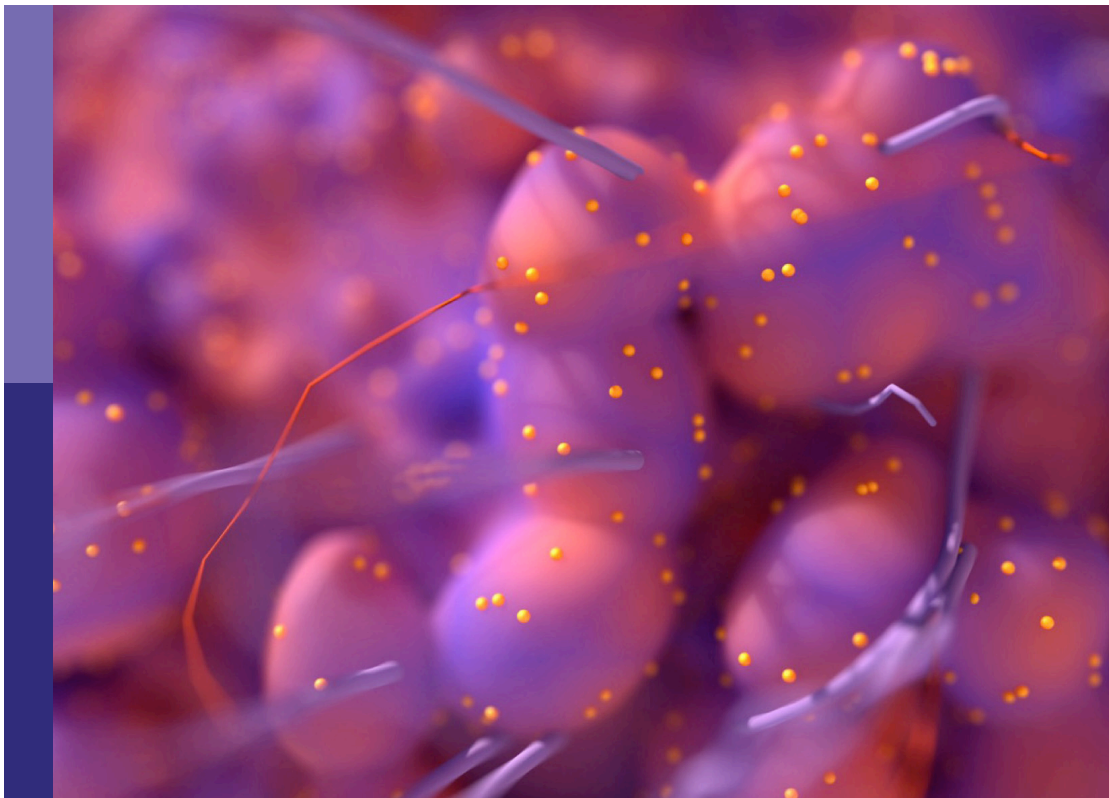
frontiers | Research Topics

The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology

Edited by
Chin-Lee Wu, Jian Lu and Xiongbing Zu

Published in
Frontiers in Oncology



ISSN 1664-8714
ISBN 978-2-83251-274-6
DOI 10.3389/978-2-83251-274-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology

Topic editors

Chin-Lee Wu — Massachusetts General Hospital, Harvard Medical School, United States

Jian Lu — Peking University Third Hospital, China

Xiongbing Zu — Central South University, China

Citation

Wu, C.-L., Lu, J., Zu, X., eds. (2023). *The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology*.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-274-6

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.

Table of contents

frontiers | Frontiers in Oncology

TYPE Editorial
PUBLISHED 21 December 2022
DOI 10.3389/fonc.2022.1118442

OPEN ACCESS

EDITED AND REVIEWED BY
Ronald M. Bukowski,
Cleveland Clinic, United States

*CORRESPONDENCE
Jian Lu
lujian@bjmu.edu.cn

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 07 December 2022
ACCEPTED 12 December 2022
PUBLISHED 21 December 2022

CITATION
Zhu X-h, Wu C-L, Zu X-b and Lu J (2022) Editorial: The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology. *Front. Oncol.* 12:1118442. doi: 10.3389/fonc.2022.1118442

Editorial: The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology

Xue-hua Zhu[1], Chin-Lee Wu[2,3], Xiong-bing Zu[4] and Jian Lu[1]*

[1]Department of Urology, Peking University Third Hospital, Beijing, China, [2]Department of Urology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States, [3]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States, [4]Department of Urology, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China

KEYWORDS

artificial intelligence (AI), machine learning, urological oncology, radiomics, clinical transformation

Editorial on the Research Topic
The application of artificial intelligence in diagnosis, treatment and prognosis in urologic oncology

Artificial intelligence (AI) is reshaping and improving traditional decision-making patterns. Although the further integration of AI into clinical practice will take time, research in this field has been producing promising results. AI has the capability to aid in several aspects of disease management. In the specialty of oncology, for example, AI can assist in disease diagnosis, classification, treatment optimization, therapeutic response evaluation, prognosis prediction, and follow-up scheduling. The greatest strength of AI is its capability to extract and interpret information from a large amount of medical data, which is included in images, pathological films, and the electronic reporting system. AI can process and analyze clinical data, but the technology cannot be fully utilized by clinicians in an advanced, efficient, and intelligent manner in order to aid in clinical decision-making.

This edition of Frontiers in Oncology seeks to assist physicians in completely comprehending and accepting AI's critical role and great potential in solving actual clinical problems, as well as to accelerate the clinical translation of relevant scientific findings. This issue consists of 10 manuscripts (including 8 original papers and 2 reviews) on the application of AI to urologic oncology and benign disease. Liu et al. develop a diagnostic tool utilizing deep learning systems (DLSs) to detect malignant cells in urine cytopathological images and then predict the histopathological results. The DLSs model achieves good predictive performance in both the internal set and extra set with an area under the curve (AUC) of 0.90 and 0.93, respectively. The model is expected to enhance the efficacy of urine biopsies and aid in the early detection and risk classification of

patients with UC. The prediction model based on a large sample database can estimate the prognosis of patients more precisely and aid in the establishment of individualized treatment plans and follow-up care. Zhang et al. created a nomogram based on the SEER database to predict cancer-specific survival (CSS) and overall survival (OS) among the elderly PCa group. This model offers precise predictions for CSS and OS in both the training and validation sets. Deng et al. develop five different machine learning (ML) models to predict the presence of prostate cancer (PCa) in patients with PSA ≤ 20 ng/mL. The random forest (RF) model performs the best among five models, with an AUC of 0.871 in the training cohort and 0.78 in the validation cohort. The research conducted by Chen et al. also focuses on the PCa prediction. Based on traditional clinical variables, five different ML models are developed and compared in terms of their predictive performance. In the test dataset, the multivariate univariate logistic regression (LR) model exhibited the best discrimination (AUC=0.918). The predictive model built by Zhang and Chen can be utilized for preliminary screening of individuals with suspected PCa, potentially reducing the number of unnecessary biopsies and missed PCa diagnoses.

Radiomics is the integration of AI with image data. The extraction and interpretation of image data expands the clinician's insight beyond the visible conventional information like lesion size, location, and shape. The acquisition of the region of interest (ROI) is a crucial step in the radiomics workflow. Manually drawing ROI is the most common method, but it is laborious and time-consuming. Chen et al. present a radiomics model based on auto-segmentation of ultrasound images to detect diabetic kidney disease (DKD) patients in three centers. DeepLabV3+ network, a DL-based anatomical-level segmentation system, demonstrates a good segmentation ability with mean pixel accuracy in the three centers of 0.890 ± 0.004, 0.870 ± 0.002 and 0.893 ± 0.007, respectively. The radiomics model delivers a fair ability for DKD recognition (AUC: 0.674 ± 0.074) and a good discriminative capability for DKD stage (AUC: 0.803 ± 0.037). Zhu et al. establish a cascaded DL model to automatically segment the whole prostate gland, the anatomic zones, and the csPCa region step by step based on biparametric MRI. The model employing cascaded convolutional neural networks (CNNs) could automatically detect and segment the suspicious csPCa lesions with excellent performance (sensitivity=95.6%, specificity=91.5%, accuracy= 92%) on MR images without any human intervention.

Ultrasound is another important imaging method for PCa diagnosis. Wang et al. propose a ML model based on 14 features extracted from transrectal ultrasound video clips of the whole prostate gland. The support vector machine (SVM) and random forest (RF) algorithms were used to establish radiomics models based on those features. The SVM model exceeds radiologist's diagnostic ability based on MRI. Imaging provides information regarding the heterogeneity and microenvironment of a tumor. Yang et al. develop and externally validate a Transformer-based DL algorithm with CT images to predict the Fuhrman nuclear grade of clear cell renal cell carcinoma (ccRCC). TransResNet, a framework network integrating CNNs, self-attention mechanisms and nonlinear classifier, outperforms conventional DL algorithms for ccRCC grade prediction with an AUC of 91.2%. This work presents a non-invasive approach for predicting the pathological grade of ccRCC, hence avoiding the complications associated with puncture biopsy.

Zhang et al. investigated the diagnostic performance of CT-radiomics in adrenal mass using a systematic review and meta-analysis. The quantitative analysis of nine studies shows that CT-based radiomics may help identify malignant adrenal tumors from benign ones with pooled sensitivity, specificity, and AUC of 0.80, 0.83, and 0.88, respectively. A mini-review by Huang et al. contains 12 articles concerning the use of radiomics in the prediction of muscle-invasive bladder cancer (BC). Radiomics utilizing CT or MRI images has the potential to detect muscle invasion.

The powerful combination of AI and medical data has yielded significant achievements in the last twenty years. Some achievements have been clinically transformed and accepted by clinicians. Studies with prospective design is required to expedite the clinical translation of AI-related findings. AI researchers should not only concentrate on the optimization of algorithms and the performance of models but also explore when and how to actualize the clinical utility of AI's achievements.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This research is supported in part by Beijing Natural Science Foundation (Z200027) and the National Natural Science Foundation of China (No. 61871004)

Acknowledgments

We thank authors of the papers published in this Research Topic for their valuable contributions and the referees for their rigorous review.

Conflict of interest

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.

Publisher's note

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 24 May 2022
doi: 10.3389/fonc.2022.901586

A Deep Learning System to Predict the Histopathological Results From Urine Cytopathological Images

Yixiao Liu[1,2,3†], Shen Jin[4†], Qi Shen[1,2,3†], Lufan Chang[5], Shancheng Fang[4], Yu Fan[1,2,3], Hao Peng[4] and Wei Yu[1,2,3*]*

[1] Department of Urology, Peking University First Hospital, Peking University, Beijing, China, [2] Institute of Urology, Peking University, Beijing, China, [3] National Urological Cancer Center, Beijing, China, [4] School of Cyber Science and Technology, Beihang University, Beijing, China, [5] R&D Department, Yizhun Medical AI Co. Ltd, Beijing, China

OPEN ACCESS

Edited by:
Jian Lu,
Peking University Third Hospital, China

Reviewed by:
Kazutoshi Fujita,
Osaka University, Japan
Satoshi Nojima,
Osaka University, Japan

***Correspondence:**
Wei Yu
yuweif@126.com
Hao Peng
penghao@act.buaa.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

Received: *22 March 2022*
Accepted: *19 April 2022*
Published: *24 May 2022*

Citation:
Liu Y, Jin S, Shen Q, Chang L, Fang S, Fan Y, Peng H and Yu W (2022) A Deep Learning System to Predict the Histopathological Results From Urine Cytopathological Images. Front. Oncol. 12:901586. doi: 10.3389/fonc.2022.901586

Background: Although deep learning systems (DLSs) have been developed to diagnose urine cytology, more evidence is required to prove if such systems can predict histopathology results as well.

Methods: We retrospectively retrieved urine cytology slides and matched histological results. High-power field panel images were annotated by a certified urological pathologist. A deep learning system was designed with a ResNet101 Faster R-CNN (faster region-based convolutional neural network). It was firstly built to spot cancer cells. Then, it was directly used to predict the likelihood of the presence of tissue malignancy.

Results: We retrieved 441 positive cases and 395 negative cases. The development involved 387 positive cases, accounting for 2,668 labeled cells, to train the DLS to spot cancer cells. The DLS was then used to predict corresponding histopathology results. In an internal test set of 85 cases, the area under the curve (AUC) was 0.90 (95%CI 0.84–0.96), and the kappa score was 0.68 (95%CI 0.52–0.84), indicating substantial agreement. The F1 score was 0.56, sensitivity was 71% (95%CI 52%–85%), and specificity was 94% (95%CI 84%–98%). In an extra test set of 333 cases, the DLS achieved 0.25 false-positive cells per image. The AUC was 0.93 (95%CI 0.90–0.95), and the kappa score was 0.58 (95%CI 0.46–0.70) indicating moderate agreement. The F1 score was 0.66, sensitivity was 67% (95%CI 54%–78%), and specificity was 92% (95%CI 88%–95%).

Conclusions: The deep learning system could predict if there was malignancy using cytocentrifuged urine cytology images. The process was explainable since the prediction of malignancy was directly based on the abnormal cells selected by the model and can be verified by examining those candidate abnormal cells in each image. Thus, this DLS was not just a tool for pathologists in cytology diagnosis. It simultaneously provided novel histopathologic insights for urologists.

Keywords: cyto-histo correlation, deep learning, urothelial carcinoma, urine cytology, convolutional neural network

INTRODUCTION

Urothelial carcinoma (UC) is one of the most common cancers worldwide (1). UCs are often multifocal and tend to recur. Thus, thorough screening and frequent surveillance are mandatory. The diagnosis of UC typically relies on the histopathological assessment of tissue resected by cystoscopy, yet it is an invasive approach and not easily accessible.

Urine cytology has played an important role in the screening and surveillance of UCs for many years for its effective, inexpensive, noninvasive nature (2–4). However, in the context of urine cytology, there is currently no gold standard for cyto-histo correlation in urine (5).

One possible solution could be to use the deep learning systems (DLSs) to build such links. DLSs have demonstrated a capacity superior to manual workflows in shifting through massive images to retrieve similar patterns and establish associations with novel traits in many medical data analysis tasks (6–8). Three studies have successfully automated the urine cytology diagnosis through the use of DLSs (9–11). One of these previous studies has further proven that DLSs might be able to determine the malignant potential of tumors more accurately than classical cytology (11). Researchers used a 16-layer convolutional neural network (CNN). Weights trained for initial UC cell detection were reused for the first 7 layers. New training with histopathological data started at the eighth layer. The DLS determined the presence of stromal invasion and performed a nuclear grading of tumor cells in the corresponding histological specimens. Therefore, DLSs have become a method that could potentially link cytopathology findings with histopathology results.

In this study, we hypothesized that routine urine cytology images contain information about the presence of malignant tissue in urinary tracts. The rationale for this cyto-histo correlation is that malignant tissues in urinary tracts undergo constant exfoliation, which sheds tumor cells and influences tumor cell morphology in urine. Building on those previous studies, we systematically investigated the presence of such correlation and aimed to capture it through our DLS. We trained and tested the DLS to spot UC cells in cytology images before using it to predict if a case would get malignant surgical pathology within the next 1 year. The results demonstrated that the DLS could predict the presence of malignancy and display such associations between cytopathology and histopathology through likelihood even without further training with histopathology data. Thus, DLS cytology can be used as not only a pathological tool to assist cytopathological diagnosis but also a novel risk-stratification tool to predict histopathology. This could help urologists make therapeutic decisions.

MATERIALS AND METHODS

Data Acquisition

All images were obtained from the archival glass of hematoxylin and eosin-stained urine cytocentrifugation cytology from consecutive patients who underwent examination, surgery, or both at Peking University First Hospital from 2014 to 2020 (**Figure 1**, **Table 1**). Urine cytology was routinely diagnosed using Papanicolaou's classification at our institute (12). Classes III, IV, and V were defined as positive; class I and II (atypical) were defined as negative.

Among the 441 positive cases (patients diagnosed with UC based on cytologic examination), 211 received surgery (surgery within the next 1 year, if not otherwise clarified), all of which were diagnosed with UC based on histological examination (**Table 2**).

Among the 395 negative cases (patients diagnosed with benign diseases based on cytologic examination), all received surgery, of which 333 were diagnosed with UC based on histological examination and the rest 62 were diagnosed with benign disease based on histological examination (**Table 2**). For the above 333 cases with contradicted cytopathological and histopathological results, a blinded pathologist's review was carried out to check for overlooked cancer cells. As a result, 63 cases actually had cancer cells in their cytology images confirmed by a pathologist and should be deemed as positive cases.

From the original slides, 1,280 × 960-pixel, Joint Photographic Experts Group (jpeg) format images were exported: 466 images from positive cases and 417 images from negative cases. Subsequently, the training–validation set and preliminary test set, internal test set, and extra test set were defined by the different cytopathological and histopathological diagnoses. The training–validation set and preliminary test set were allocated by 8:1 stochastically. This study was approved by the institutional review board of Peking University First Hospital.

Deep Learning System

We built our DLS on a ResNet101 Faster-RCNN (**Figure 2**). ResNet101 is a 101-layer Reidual Network proposed on the 2016 IEEE Conference on Computer Vision and Pattern Recognition by He et al. (13). Deeper and restructured, ResNet101 has shown a high performance in many contexts of use including skin lesion detection and brain disease detection in magnetic resonance images (14, 15).

Faster-RCNN, short for faster region-based convolutional neural network, is a CNN that combines object detection and classification into one network (16). It extracts features, makes detections through these features, and quantifies the degree of fit at each detection using a value of possibility ranging from the worst of 0 to the optimum of 100. Evidence has shown that Faster-RCNN is especially good at detecting objects at multiple scales and aspect ratios, such as abnormal cervical cells in cytology images and cancer regions in colorectal biopsies (17, 18).

The model was implemented in Python 3.8 using TensorFlow (1.12.0) and Keras (2.0.3). Malignant cells with remarkable atypia in the jpeg images were annotated by a certified urological pathologist using the open-source software LabelMe (19). The images were then divided into 175 × 200-pixel panel subimages automatically, which were used for the training of the DLS.

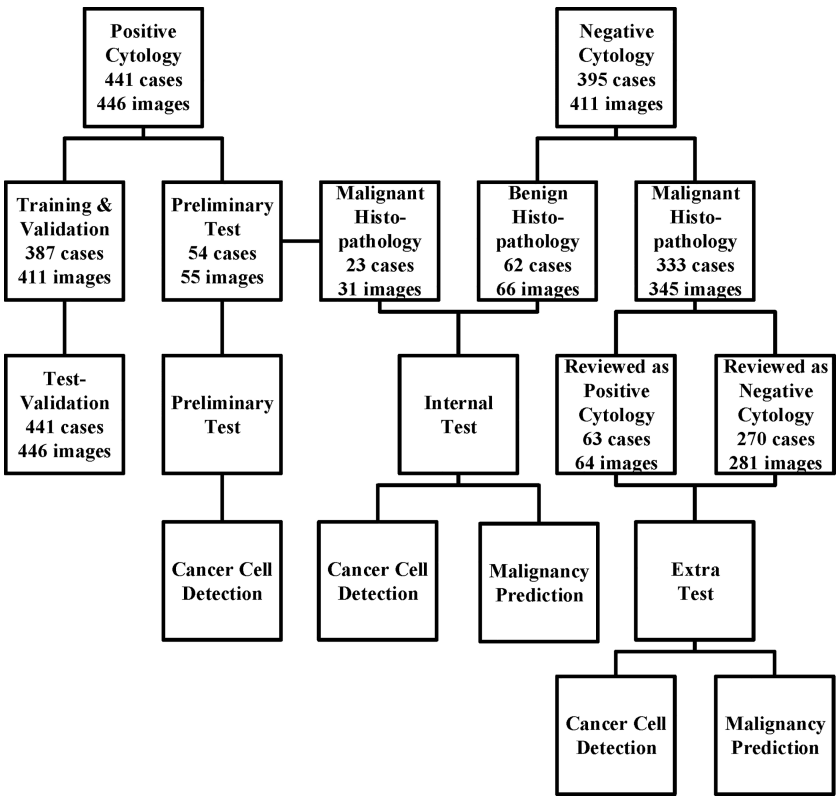


FIGURE 1 | The data acquisition process is illustrated here. We collected the cytology images retrospectively and consecutively from Sep 2014 to Jan 2020 and built a series of data sets for training, validation, and tests. For those who underwent surgeries within the next 1 year, surgical results were also followed. The preliminary test set was only used for cancer cell detection, while the internal and extra test sets were used for both cancer cell detection and malignancy prediction.

ResNet101 was pretrained on the ImageNet database consisting of 1.2 million training images, with 1,000 classes of objects (20). The weight pretrained with ImageNet was used to initiate the weights of all convolutional layers, and all weights were trained with cytology images afterwards. The images passed through 33 convolution blocks and then through 1 dense layer. The SoftMax function was used as the activation function.

TABLE 1 | Baseline characteristics.

	Training and Validation	Preliminary Test	Negative Cytology with Benign Histopathology	Negative Cytology with Malignant Histopathology
Age				
<60	79	10	26	26
≥60	308	44	36	307
Sex				
Female	260	11	23	95
Male	127	43	43	238
Cytology diagnosis[1]				
I	0	0	57	248
II	0	0	5	85
III	260	40	0	0
IV	127	14	0	0
V	0	0	0	0

[1]*Cytology was diagnosed following Papanicolaou classification.*

The training set and validation set were further allocated by 5:1 stochastically, for early stopping during network training to suppress overfitting. Spatial augmentation, including 90° rotation and vertical and horizontal flip, was applied in network training. We set 80 as the maximum epoch and stopped training if validation loss did not improve after 15 epochs.

For the prediction of malignancy, an additional classifier was added at the end of the initial DLS. The function of the classifier was to select the highest value of the possibilities in an image and to make a binary classification (benign or malignant) by comparing this value to the threshold (**Figure 2**). More details were provided in the **Supporting Materials (Supporting File 1)**.

Evaluation Metrics

Performance was evaluated based on the testing results.

For the detection of UC cells, the annotations served as the reference standard. We used sensitivity, accuracy, and average false-positive cells per image. Sensitivities and accuracies were calculated using following formulas:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{Total Annotations}} \quad ; \quad \text{Accuracy} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

For the prediction of the malignancy, the surgical results served as the reference standard. We used sensitivity, specificity, the F1 score, and kappa score for evaluation. Cohen kappa

TABLE 2 | Surgical follow-up.

	Training and Validation Set[1]	Positive Cytology with Malignant Histology	Negative Cytology with Benign Histology	Extra Test Set
N	188	23	62	333
Sex				
Female	63	5	19	95
Male	125	18	43	238
Age				
<60	26	3	26	74
≥60	162	20	36	259
Surgery[2]				
TUR-Bt or biopsy	102	16	48	310
nephroureterectomy	100	6	13	34
Radical cystectomy	9	1	1	20
Tumor				
Negative	0	0	62	0
upper urinary tract	89	6	0	15
Lower urinary tract	92	17	0	301
Synchronous U&L	7	0	0	17
tumor grade				
Low grade	24	7		117
High grade	164	16		213
NA[3]	0	0		3
Tumor stage				
Muscle non-invasive	104	16		277
Muscle invasive	82	6		52
NA[4]	2	1		4

Cytology was diagnosed following Papanicolaou criteria. Cancer grade was diagnosed using WHO2004. Tumor stage was diagnosed using TNM staging AJCC UICC 8th edition. Synchronous U&L, synchronous tumors in both upper and lower urinary tract. NA, not available.

[1]Only cases in the training and validation set who underwent surgery were listed here.

[2]Many cases undertook more than 1 procedure, either at one time or many times.

[3]Tumor grades were missing due to the following: a case reported as unable to rule out for low grade (n = 1); grade not reported for a case with in situ carcinoma (n = 1); a case reported as Grade 2 using WHO 1999 but not using WHO 2004 (n=1).

[4]Tumor stages were missing for those undertaken biopsies with no further operation available (n = 2 + 1 + 4).

scores reflect the agreement of the DLS with the pathologist reference standard (21). F1 scores were calculated using the following formula:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

RESULTS

Development of the Deep Learning System

The first step of our model is to detect just cancer cells. Therefore, we only annotated the cancer cells in the jpeg images for training and validation purposes, and treated all the other cells in the same image as background (**Supporting File1**). In total, 1,364 cells from 411 1,280×960-pixel images were labelled. A total of 1953 subimages were obtained. Each subimage contained at least one label in it. Sub-images were subsequently randomly allocated into the training and validation sets by the ratio of 5:1, and subjected to the pretrained DLS, as mentioned in *Materials and Methods*. Both the total loss and the system accuracy stabilized after 45–50 epochs for the validation set. A final model was chosen at 48 epochs when the total loss for the validation set hit the lowest point of 1.6. It was also where the classification accuracy for the validation set hit the highest point of 0.77.

Deep Learning System Performance to Detect Cancer Cells

We evaluated the ability of the DLS to detect cancer cells in the three test sets (**Figure 3**). We took advantage of the value of possibility generated by the DLS and adopted it as the threshold for cell detection.

For all sets, sensitivities increased at the cost of more cells mistakenly spotted as malignant by the DLS (**Figure 4A**). The accuracy initially increased with the thresholds for cell detection, and the rate of increase slowed down by approximately 50–55 (**Figure 4B**). Such a trend was observed in both preliminary and internal tests. Therefore, we chose 55 as the optimal threshold. Under this threshold, the sensitivity is 41% for the preliminary test at the cost of an average of 3.09 false-positive cells per image, 36% for the internal test at the cost of an average of 0.72 false-positive cells per image, and 41% for the extra test at the cost of an average of 0.31 false-positives cell per image (**Figure 4A**). The accuracy of cancer cell detection was 50.0, 50.3, and 14.5 for the preliminary test, internal test, and extra test, respectively (**Figure 4B**).

For the subgroups of the extra test set, results are also calculated under the optimal threshold of 55: for the positive 64 images, sensitivity is 41% with an average of 0.95 false-positive cells per image; for the 281 negative images, the average false-positive cell per image is 0.16 while the sensitivity is not available due to no true positive (cancer cell).

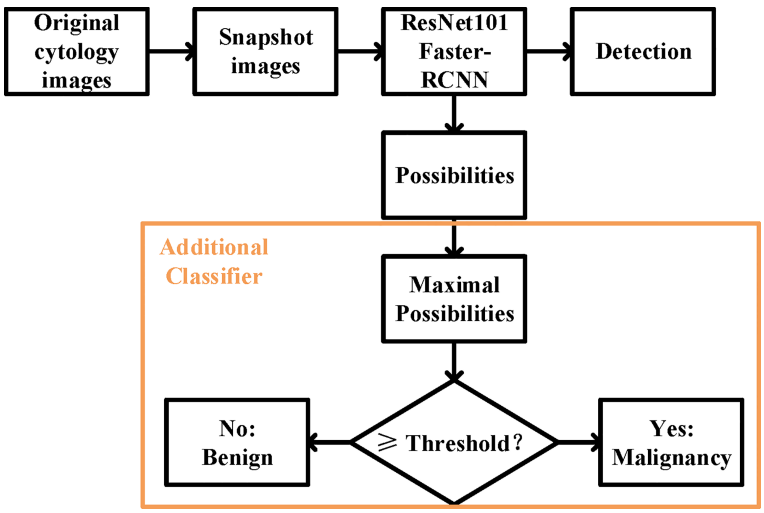


FIGURE 2 | The overall design of the deep learning system. The ResNet 101 Faster-RCNN detected the UC cells while assigning a possibility to each of the cell, and an additional classifier picked the maximal possibility and predicted the histopathological malignant state according to the set threshold.

Deep Learning System Performance to Predict Malignancy

We evaluated whether the exact same DLS can predict malignancy with no further transfer learning. Because the gold standard of malignancy is histopathology, we paired cytology with its corresponding surgical pathology. A total of 97 1,280×960-pixel images in the internal test set and 345 1,280×960-pixel images in the extra test set that could be paired with corresponding histopathological specimens were used. Here, we proposed a hypothesis that a case with positive cytology was more likely to have malignant surgical pathology. Therefore, the maximal value of possibility (the threshold for cell detection) in each image was adopted as the threshold for malignancy prediction.

Notably, the DLS was able to predict malignancy through cytology images. For the internal test, the AUC was 0.90 (95%CI 0.84–0.96) (**Figure 5A**). The highest kappa score is 0.71 at the threshold of 57 for malignancy prediction, and the highest F1 score is 0.78 at the threshold of 58 (**Table 3**). For the extra test,

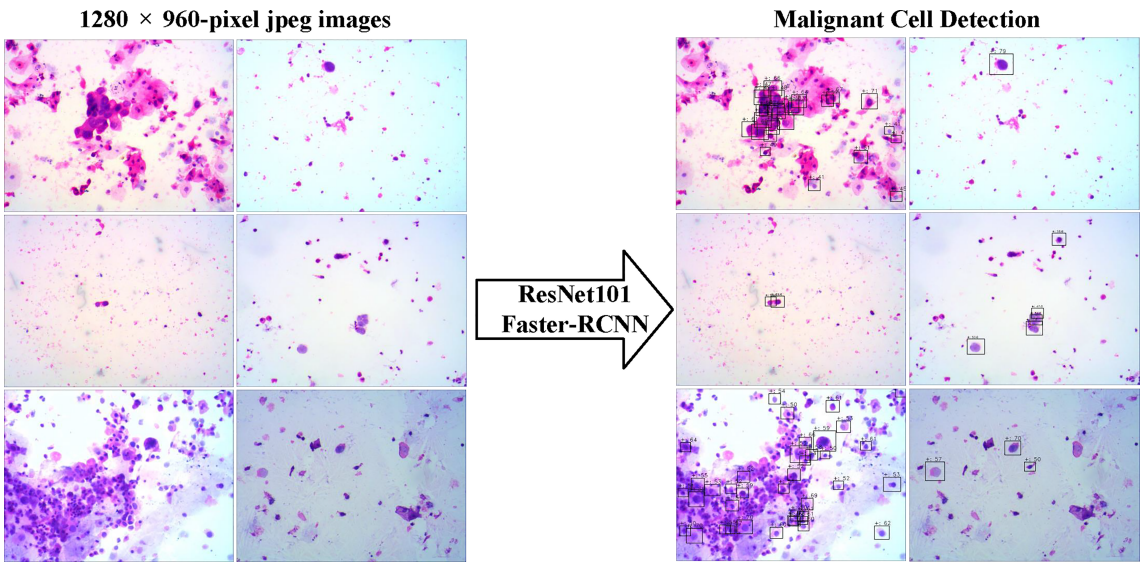


FIGURE 3 | The examples of snapshot images from positive cases and the results by the deep learning system were provided. The malignant cells detected were labeled by the Faster-RCNN, and the possibilities of each detection were also shown.

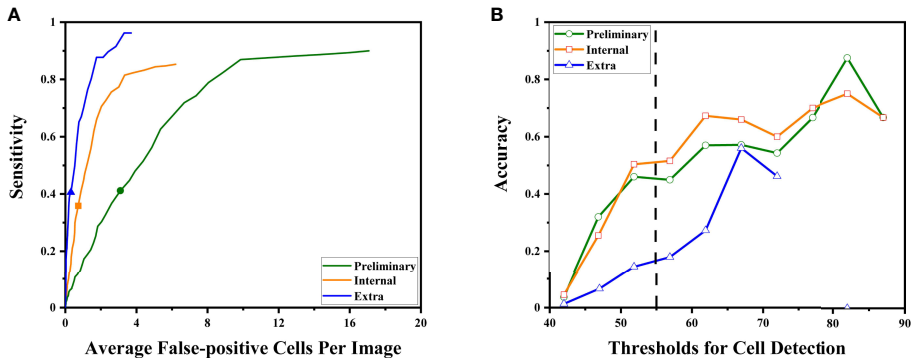


FIGURE 4 | The DLS performance to detect malignant cells is illustrated. **(A)** As the system tried to find more malignant cells (to achieve higher sensitivity), it made more mistakes classifying benign cells as malignant. The sensitivities at the optimal threshold for three sets were marked. **(B)** The accuracy at different thresholds. Thresholds were represented by their middle value (for example, x = 52.5 represented the interval of 50–55 points). Higher thresholds tend to have better detection accuracy. The performance at the optimal threshold of 50–55 points is highlighted by the dashed line. At this point, the increasing rates began to slow down.

the AUC was 0.93 (95%CI 0.90–0.95) (**Figure 5A**). The highest kappa score is 0.60 at the threshold of 52 for malignancy prediction, and the highest F1 score is 0.69 at the threshold of 52 (**Table 4**).

Under the optimal threshold (threshold=55), the DLS also achieved good performance. For the internal test, 4 images that scored higher than 55 have benign histologic results (**Figure 5B**). Thus, sensitivity is 71% (95%CI 52%–85%); specificity is 94% (95% CI 84%–98%); the F1 score is 0.76; and there was a substantial agreement with the reference standard (kappa = 0.68 [95%CI 0.52–0.84]). For the extra test, 22 images that scored higher than 55 have benign histologic results (**Figure 5C**). Thus, sensitivity is 67% (95% CI 54%–78%); specificity is 92% (95%CI 88%–95%); the F1 score is 0.66; and there was a moderate agreement with the reference standard (kappa = 0.58 [95%CI 0.46–0.70]).

DISCUSSION

In this study, we developed a DLS to predict the likelihood of the presence of tissue malignancy through urine cytopathology. Notably, the system achieved an AUC of 0.90 for the internal test and of 0.93 for the extra test. Under the optimal threshold, sensitivity is 71%, and specificity is 94% for the internal test; sensitivity is 67%, and specificity is 92% for the extra test. These results proved that the DLS was able to predict the presence of malignant tissue merely from urine cytology images.

It has been fully demonstrated that deep learning models can be used to establish a cytology diagnosis system. Vaickus et al. were the first to show that the analysis of urine cytology specimens could be reliably automated. They achieved an accuracy of more than 90% using a hybrid deep-learning and morphometric algorithm (9). Pantanowitz et al. further proved this idea using a much larger data set. They used a pure neural network to exploit and integrate both slide-level and cell-level features and achieved a sensitivity of 79.5% and a specificity of 84.5% for cytopathological diagnosis (10). For both studies, features were carefully engineered to ensure biological interpretability and reproducibility. Features such as the nuclear-cytoplasm ratio, chromatin quality, and the quantity of cells were included (**Figure 6**). Such a design made the system explainable and enables it to fulfill the aim to assist pathologists in cytology reading.

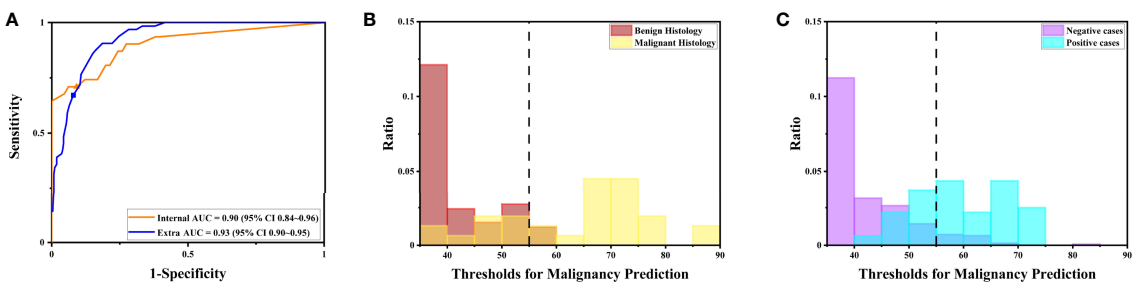


FIGURE 5 | The DLS performance to predict malignancy is illustrated. **(A)** The receiver operating characteristics curve for the performance to predict malignancy for the internal and extra tests. **(B)** The distribution of maximal possibilities in the internal test set. Most images from benign histology were scored below 55. **(C)** The distribution of maximal possibilities in the extra test set. Most images from negative cases were also scored below 55 (highlighted by the dashed line).

TABLE 3 | Performance on the internal test set.

Threshold	Sensitivity	Specificity	F1 Score	Kappa Score
40.00	0.94	0.61	0.67	0.45
42.00	0.94	0.62	0.67	0.51
43.00	0.90	0.68	0.70	0.54
44.00	0.90	0.71	0.72	0.56
45.00	0.90	0.73	0.71	0.55
46.00	0.87	0.74	0.72	0.57
47.00	0.87	0.76	0.73	0.57
49.00	0.81	0.79	0.71	0.58
50.00	0.81	0.80	0.69	0.56
51.00	0.74	0.83	0.71	0.56
53.00	0.74	0.88	0.72	0.64
54.00	0.71	0.91	0.75	0.68
55.00	0.71	0.94	0.77	0.68
57.00	0.68	0.95	0.76	0.71
60.00	0.65	1.00	0.78	0.71
65.00	0.61	1.00	0.76	0.68
68.00	0.48	1.00	0.59	0.50
69.00	0.42	1.00	0.56	0.46
70.00	0.39	1.00	0.41	0.32
71.00	0.26	1.00	0.32	0.25
73.00	0.19	1.00	0.32	0.25
75.00	0.16	1.00	0.23	0.17
76.00	0.13	1.00	0.23	0.17
78.00	0.10	1.00	0.18	0.13
83.00	0.06	1.00	0.12	0.09
88.00	0.00	1.00	NA	NA

NA, not available.

TABLE 4 | Performance on the extra test set.

Threshold	Sensitivity	Specificity	F1 Score	Kappa Score
40.00	1.00	0.56	0.52	0.35
41.00	1.00	0.59	0.54	0.37
42.00	0.98	0.62	0.57	0.42
43.00	0.98	0.67	0.58	0.44
44.00	0.97	0.69	0.60	0.47
45.00	0.97	0.72	0.62	0.50
46.00	0.94	0.75	0.63	0.51
47.00	0.91	0.78	0.67	0.56
48.00	0.91	0.81	0.68	0.58
49.00	0.88	0.84	0.68	0.59
50.00	0.86	0.85	0.68	0.60
51.00	0.83	0.86	0.68	0.60
52.00	0.80	0.88	0.69	0.60
53.00	0.77	0.89	0.66	0.58
54.00	0.72	0.90	0.67	0.59
55.00	0.67	0.92	0.66	0.58
56.00	0.63	0.94	0.64	0.57
57.00	0.59	0.94	0.61	0.54
58.00	0.55	0.95	0.58	0.51
59.00	0.48	0.96	0.55	0.48
60.00	0.45	0.96	0.53	0.45
61.00	0.42	0.96	0.52	0.45
62.00	0.41	0.96	0.53	0.47
63.00	0.39	0.98	0.50	0.44
64.00	0.36	0.98	0.49	0.44
65.00	0.34	0.99	0.47	0.41
66.00	0.31	0.99	0.37	0.32
67.00	0.23	0.99	0.24	0.20
68.00	0.14	1.00	0.24	0.20
70.00	0.13	1.00	0.19	0.16
71.00	0.11	1.00	0.14	0.11
72.00	0.08	1.00	0.09	0.07
73.00	0.05	1.00	0.06	0.04
74.00	0.03	1.00	NA	NA
78.00	0.00	1.00	NA	NA

NA, not available.

Although the principal indications for the use of cytology include the diagnosis, follow-up, and monitoring of patients with urothelial tumors, the gold standard for the diagnosis of tumor is still histopathology. Therefore, the cyto-histo correlation has become the bottleneck for the use of cytology in clinical settings. Previous studies have shown that deep learning techniques could discern subtle differences in image features that are not readily noticeable to pathologists between tissues from patients with different genetic subtypes, cancer grades, and survival (6–8). Thus, it was reasonable to wonder if deep learning could also distinguish differences in cytology images with different histopathological results.

Fujita et al. were the first to observe that the DLS could not only accurately detect UC cells but also distinguish characteristics traditionally determined using histopathology. They probed two specific characters: whether the lesions were invasive and whether the lesions were high grade. For both, the DLS achieved an AUC higher than 0.86 and F1 score higher than 0.82 (11). The study used a mixing-training model with the first half trained with cytopathological data and the other half trained with histopathological data (**Figure 6**). The results proved that at least part of the features could be shared for cytopathology diagnosis and histopathology prediction. This laid the foundation for a complete cyto-histo correlation.

For the purpose of proving this correlation more stringently, we improved the design of this study on the basis of previous ones. First, we trained the system with only positive cytology images instead of adding negative images as in previous studies. Therefore, this design could insure all the features used by the DLS derived from malignant cases. Second, during malignancy prediction, we did not train the initial DLS again. Instead, we added an additional classifier at the end of the initial DLS. This design rendered two models that shared a same set of convolutional network and detector and, thus, a same set of features. The additional classifier was designed based on the biological meaning of the degree of fit calculated by the DLS. During cancer cell detection, the degree of fit was the likelihood of a cell to be a cancer cell. Therefore, the maximal degree of fit in a certain image represented the likelihood of a case to get malignancy. There were no manually designed features in the DLS. Therefore, the process was explainable since the prediction of malignancy was directly based on the abnormal cells selected by the model and can be verified by examining those candidate abnormal cells in each image. Thus, the design made the DLS not only possible to testify the hypothesis better but also explainable in a unique way.

The test sets in this study included preliminary, internal, and extra test sets. The extra test set came from patients whose cytology was initially diagnosed as negative but later proved to have malignant histopathological results. Overlooked cancer cells

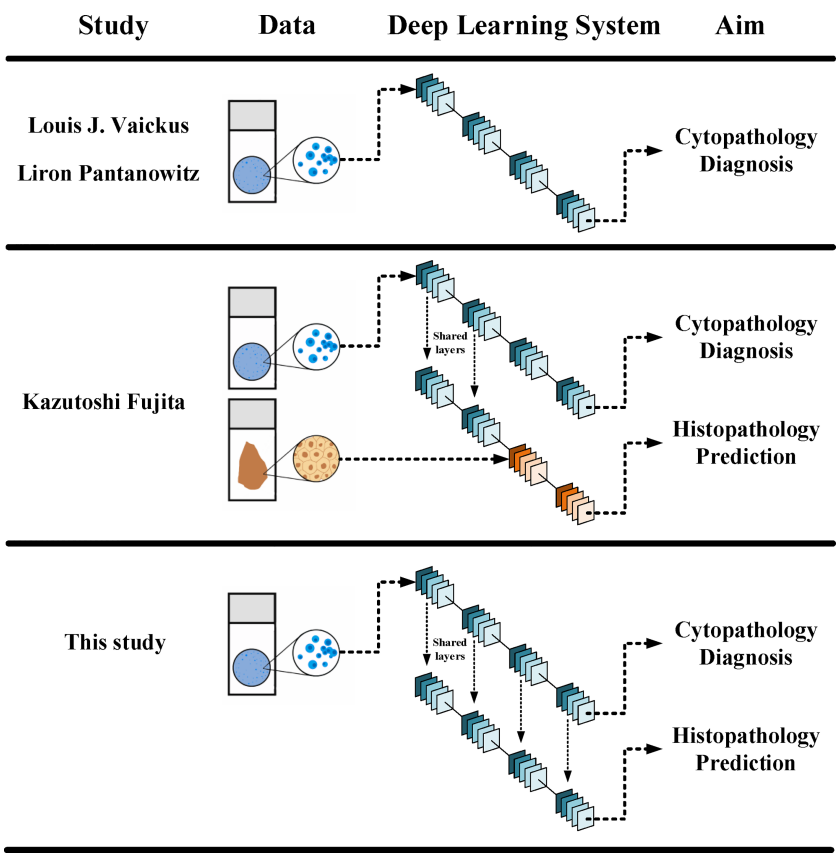


FIGURE 6 | Different designs in between different DLSs. The current DLS could make both cytopathological diagnosis and histopathological prediction, yet it did not need histopathological data in its training.

in these images were carefully revised in the blinded review at the beginning of this study. Results showed that the extra test set had higher AUC than the internal test set during malignancy prediction. This indicated that the DLS had a good performance for the extra test despite the fact that these cases were among the most difficult to diagnose by traditional cytology. However, the sensitivity and specificity for the extra test under the optimal threshold were lower than that for the internal test. Moreover, during cancer cell detection, the accuracy for the extra test was also lower than that for the preliminary and the internal tests under the optimal threshold, and the accuracy for the extra test was still increasing after the threshold reached the optimal threshold. This indicated that the current optimal threshold using in this study, which was chosen based on the preliminary and internal tests, might not serve as the optimal threshold for the extra test. Future studies are needed for a better strategy to find the optimal threshold for DLS application. At the same time, it is also important to understand that pathologists are not able to spot every true cancer cell since no cytopathological scoring system used for cytology diagnosis at present is perfect. A DLS learnt from pathologists' annotations inevitably inherited these bias and errors. Therefore, as pathologists had failed to perform excellent in the extra test set themselves, the DLS would only detect cancer cells with additional difficulties.

The results of the DLS performance to predict malignancy showed a relatively high specificity. The sensitivity, however, was not as good as the specificity. This indicated that most cases predicted to have tissue malignancy were indeed patients with UC, while some patients with UC were not successfully identified. This may be attributed to the fact that the DLS still needs further improvement or that some UCs do not present morphologically abnormal cells in urine. Meanwhile, it was not necessary to spot every cancer cell to make a prediction. Instead, it was adequate to find most of the cancer cells while mistaking as few normal cells as possible. These successfully spotted cancer cells were likely to be among the most atypical and, thus, gave the highest scores in an image. Biologically, it was also the patients whose cells in cytology had a higher degree of atypia who were more likely to get UC. Future studies are needed to identify UC in those patients without abnormal cells in urine.

There is currently no gold standard for cyto-histo correlation in urine. Many argue that a negative cytology with a concurrent positive surgical result is not a false negative. Similarly, a positive urine followed by a negative surgical result is not a false positive.

However, results in this current study imply that when the DLS predicts a malignant state, it might focus on characteristics that are partially same with those used for UC cell detection. This indicates that there are features on cytology images correlated to histopathology results. Nevertheless, due to the lack of technical maneuver to untangle representative features in Faster-RCNN, we are not able to define each feature and apply them in classical cytology. This is one of the limitations of this study. Another limitation is that this is a retrospective study in a single center. Multicentered prospective studies are warranted to further prove these findings.

Collectively, the current results demonstrated that DLS cytology could be used to predict the likelihood of a case to have histological confirmed malignancy through a cyto-histo correlation. If a DLS can serve as a risk-stratification tool to distinguish clinically relevant malignancy at the time of cytology, urologists can plan in time therapeutic strategies at lower cost that benefit more patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of biological and medical research of Peking University First Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YL, SJ, and WY set up the experimental design. YL, QS, and YF collected and selected the data. SJ was mainly responsible for the high-performance computing and statistical analysis. QS was responsible for annotations and the morphology interpretation in the cytology and histology. LC provided the basic algorithm model. LC and SF provided guidance on deep learning and computer vision techniques. YL wrote the manuscript. WY and HP were responsible for conception, and they supervised the work. All authors reviewed the manuscript and agree with its contents.

FUNDING

This study was funded by National Natural Science Foundation of China (81870518 to WY and 62073012 to HP).

ACKNOWLEDGMENTS

We are grateful to the patients and urological pathologists at the Peking University First Hospital. YG, HH, and YM are acknowledged for their advice on the setup of the model. XY and LZ are acknowledged for their advice and guidance on the illustration of figures. We express our sincere appreciation for the grants from the National Natural Science Foundation of China (81870518 to WY and 62073012 to HP).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.901586/full#supplementary-material

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
2. Witjes JA, Bruins HM, Cathomas R, Comperat EM, Cowan NC, Gakis G, et al. European Association of Urology Guidelines on Muscle-Invasive and Metastatic Bladder Cancer: Summary of the 2020 Guidelines. *Eur Urol* (2020) 79(1):82–104. doi: 10.1016/j.eururo.2020.03.055
3. Flaig TW, Spiess PE, Agarwal N, Bangs R, Boorjian SA, Buyyounouski MK, et al. Bladder Cancer, Version 3.2020, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* (2020) 18(3):329–54. doi: 10.6004/jnccn.2020.0011
4. Roupret M, Babjuk M, Burger M, Capoun O, Cohen D, Comperat EM, et al. European Association of Urology Guidelines on Upper Urinary Tract Urothelial Carcinoma: 2020 Update. *Eur Urol* (2021) 79(1):62-79. doi: 10.1016/j.eururo.2020.05.042
5. Crothers BA. Cytologic-Histologic Correlation: Where Are We Now, and Where Are We Going? *Cancer Cytopathol* (2018) 126(5):301–8. doi: 10.1002/cncy.21991
6. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-Cancer Image-Based Detection of Clinically Actionable Genetic Alterations. *Nat Canc* (2020) 1(8):789–99. doi: 10.1038/s43018-020-0087-6
7. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis. *Nat Canc* (2020) 1(8):800–10. doi: 10.1038/s43018-020-0085-8
8. Mi H, Bivalacqua TJ, Kates M, Seiler R, Black PC, Popel AS, et al. Predictive Models of Response to Neoadjuvant Chemotherapy in Muscle-Invasive Bladder Cancer Using Nuclear Morphology and Tissue Architecture. *Cell Rep Med* (2021) 2(9):100382. doi: 10.1016/j.xcrm.2021.100382
9. Vaickus LJ, Suriawinata AA, Wei JW, Liu X. Automating the Paris System for Urine Cytopathology-A Hybrid Deep-Learning and Morphometric Approach. *Cancer Cytopathol* (2019) 127(2):98–115. doi: 10.1002/cncy.22099
10. Sanghvi AB, Allen EZ, Callenberg KM, Pantanowitz L. Performance of an Artificial Intelligence Algorithm for Reporting Urine Cytopathology. *Cancer Cytopathol* (2019) 127(10):658–66. doi: 10.1002/cncy.22176
11. Nojima S, Terayama K, Shimoura S, Hijiki S, Nonomura N, Morii E, et al. A Deep Learning System to Diagnose the Malignant Potential of Urothelial Carcinoma Cells in Cytology Specimens. *Cancer Cytopathol* (2021) 129 (12):984–95. doi: 10.1002/cncy.22443
12. Owens CL, Vandenbussche CJ, Burroughs FH, Rosenthal DL. A Review of Reporting Systems and Terminology for Urine Cytology. *Cancer Cytopathol* (2013) 121(1):9–14. doi: 10.1002/cncy.21253
13. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition (2016) (Accessed 2016 IEEE Conference on Computer Vision and Pattern Recognition).

14. Dai D, Dong C, Xu S, Yan Q, Li Z, Zhang C, et al. Ms RED: A Novel Multi-Scale Residual Encoding and Decoding Network for Skin Lesion Segmentation. *Med Imag Anal* (2022) 75:102293. doi: 10.1016/j.media.2021.102293
15. Chen H, Dou Q, Yu L, Qin J, Heng PA. VoxResNet: Deep Voxelwise Residual Networks for Brain Segmentation From 3D MR Images. *Neuroimage.* (2018) 170:446–55. doi: 10.1016/j.neuroimage.2017.04.041
16. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* (2017) 39(6):1137–49. doi: 10.1109/TPAMI.2016.2577031
17. Li X, Xu Z, Shen X, Zhou Y, Xiao B, Li TQ. Detection of Cervical Cancer Cells in Whole Slide Images Using Deformable and Global Context Aware Faster RCNN-FPN. *Curr Oncol* (2021) 28(5):3585–601. doi: 10.3390/curroncol28050307
18. Ho C, Zhao Z, Chen XF, Sauer J, Saraf SA, Jialdasani R, et al. A Promising Deep Learning-Assistive Algorithm for Histopathological Screening of Colorectal Cancer. *Sci Rep* (2022) 12(1):2222. doi: 10.1038/s41598-022-06264-x
19. Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: A Database And Web-Based Tool for Image Annotation. *Int J Of Comput Vision* (2008) 77(1-3):157–73. doi: 10.1007/s11263-007-0090-8
20. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* (2015) 115 (3):211–52. doi: 10.1007/s11263-015-0816-y
21. Fleiss JL, Cohen J, Everitt BS. Large Sample Standard Errors of Kappa And Weighted Kappa. *Psycholog Bullet* (1969) 72:323–7. doi: 10.1037/h0028106

Conflict of Interest: Authors YL, SJ, QS, LC, SF, YF, HP, and WY are on a pending patent (China 202110493539.5). Author LC was employed by the company Yizhun Medical AI Co. Ltd.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fonc.2022.918780

Development and Validation of Nomograms to Predict Cancer-Specific Survival and Overall Survival in Elderly Patients With Prostate Cancer: A Population-Based Study

Zhaoxia Zhang[1,2,3], *Chenghao Zhanghuang*[1,2,3,4], *Jinkui Wang*[1,2,3], *Xiaomao Tian*[1,2,3], *Xin Wu*[1,2,3], *Maoxian Li*[1,2,3], *Tao Mi*[1,2,3], *Jiayan Liu*[1,2,3], *Liming Jin*[1,2,3], *Mujie Li*[1,2,3] *and Dawei He*[1,2,3*]

[1] Department of Urology, Chongqing Key Laboratory of Children Urogenital Development and Tissue Engineering, Chongqing, China, [2] Chongqing Key Laboratory of Pediatrics, Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China, [3] National Clinical Research Center for Child Health and Disorders, China International Science and Technology Cooperation base of Child development and Critical Disorders; Children's Hospital of Chongqing Medical University, Chongqing, China, [4] Department of Urology, Kunming Children's Hospital, Yunnan Provincial Key Research Laboratory of Pediatric Major Diseases, Kunming, China

OPEN ACCESS

Edited by:
Xiongbing Zu,
Central South University, China

Reviewed by:
Maolake Aerken,
University at Buffalo, United States
Rongjun Wan,
Central South University, China

***Correspondence:**
Dawei He
hedawei@hospital.cqmu.edu.cn

Specialty section:
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

Received: *12 April 2022*
Accepted: *23 May 2022*
Published: *23 June 2022*

Citation:
Zhang Z, Zhanghuang C, Wang J, Tian X, Wu X, Li M, Mi T, Liu J, Jin L, Li M and He D (2022) Development and Validation of Nomograms to Predict Cancer-Specific Survival and Overall Survival in Elderly Patients With Prostate Cancer: A Population-Based Study. Front. Oncol. 12:918780. doi: 10.3389/fonc.2022.918780

Objective: Prostate cancer (PC) is the most common non-cutaneous malignancy in men worldwide. Accurate predicting the survival of elderly PC patients can help reduce mortality in patients. We aimed to construct nomograms to predict cancer-specific survival (CSS) and overall survival (OS) in elderly PC patients.

Methods: Information on PC patients aged 65 years and older was downloaded from the Surveillance, Epidemiology, and End Results (SEER) database. Univariate and multivariate Cox regression models were used to determine independent risk factors for PC patients. Nomograms were developed to predict the CSS and OS of elderly PC patients based on a multivariate Cox regression model. The accuracy and discrimination of the prediction model were tested by the consistency index (C-index), the area under the subject operating characteristic curve (AUC), and the calibration curve. Decision curve analysis (DCA) was used to test the clinical value of the nomograms compared with the TNM staging system and D'Amico risk stratification system.

Results: 135183 elderly PC patients in 2010-2018 were included. All patients were randomly assigned to the training set (N=94764) and the validation set (N=40419). Univariate and multivariate Cox regression model analysis revealed that age, race, marriage, histological grade, TNM stage, surgery, chemotherapy, radiotherapy, biopsy Gleason score (GS), and prostate-specific antigen (PSA) were independent risk factors for predicting CSS and OS in elderly patients with PC. The C-index of the training set and the validation set for predicting CSS was 0.883(95%CI:0.877-0.889) and 0.887(95% CI:0.877-0.897), respectively. The C-index of the training set and the validation set for predicting OS was 0.77(95%CI:0.766-0.774)and 0.767(95%CI:0.759-0.775),

respectively. It showed that the proposed model has excellent discriminative ability. The AUC and the calibration curves also showed good accuracy and discriminability. The DCA showed that the nomograms for CSS and OS have good clinical potential value.

Conclusions: We developed new nomograms to predict CSS and OS in elderly PC patients. The models have been internally validated with good accuracy and reliability and can help doctors and patients to make better clinical decisions.

Keywords: nomogram, old age, prostate cancer, CSS, SEER

BACKGROUND

Prostate cancer (PC) is the most common non-dermatological tumor in men worldwide. In 2022, the number of new prostate cancer patients in the United States will reach 268,490 (1). Most prostate cancer can be diagnosed early due to the popularity of prostate-specific antigen (PSA) screening and biopsy testing techniques. The treatment of prostate cancer mainly includes radical prostatectomy, androgen deprivation therapy (ADT), radiotherapy, and chemotherapy, which dramatically improves prostate cancer patients' survival rate. At the same time, the total number of prostate cancer patients is also increasing. As of 2020, there were 3.65 million confirmed PC patients diagnosed in the United States, and the number is expected to increase to 5.02 million by 2030 (2). It should be noted that mortality rates among prostate cancer patients also rank second among cancer deaths worldwide due to their high morbidity. Although most patients with PC have a good prognosis, some patients still have a recurrence and distant metastasis, making significant differences in the prognosis of PC. In 2022,34,500 people are expected to die from PC in the United States or about 11% of male cancer deaths (1).

Previously, the US Joint Commission on Cancer (AJCC) tumor-lymph node-metastatic (TNM) cancer staging system was used for the effective management of a variety of cancers (3). However, a growing number of studies have shown that the TNM stage alone does not accurately predict patient outcomes because of multiple factors clinically associated with PC prognosis (4–6), especially the Gleason score(GS) and prostate-specific antigen (PSA). GS is the most powerful tool for predicting outcomes of PC (7), developed by Donald Gleason Joint Urology Research Group between 1966 and 1974 (8) and revised in 2005 and 2014 (9, 10). PSA is mainly used for the screening of PC, causing a significant increase in the detection rate of PC. It is well known that PSA level is an essential factor in determining the aggressiveness of prostate cancer (11), and some studies show that PSA level is considered an essential prognostic factor in PC, with a linear relationship between PSA and PC prognosis (12, 13). However, some studies have shown that PSA screening does not reduce all-cause mortality in patients with PC (14, 15). Although the combination of PSA, histological grade, and TNM staging system can establish prognostic models, refining the stratification system can improve the discriminatory ability of prognostic models (16, 17). However, this model still cannot evaluate the impact of critical clinical variables, including age, marriage, race, and treatment mode, on the prognosis of PC patients.

A nomogram is a digital graphical tool that can predict the occurrence probability of a given event based on the data of known variables. It is considered superior to the conventional TNM staging system (18, 19). It has been widely used to predict the prognosis of multiple cancers, including glioma, bladder cancer, renal carcinoma, mammary cancer, and Colon cancer (20–24). There are also some nomograms for PC, but primarily for distant metastatic PC, PC with particular bone metastases, or patients with nonmetastatic PC. There are also nomograms designed specifically for PC patients with GS 3 + 4 and 4 + 3 scores (25–28).Elderly patients are a group with high incidence and mortality from PC, with a median age at diagnosis of 66 years (29). Moreover, with the aging population, the base of the elderly is also expanding, and the cancer health management of the elderly has become a major problem that cannot be ignored (30). More than 60 percent of PC patients are over the age of 65, and more than 90 percent of PC deaths occur in this age group (6), but there is no nomogram for PC patients more than 65 years old. Consider that elderly PC patients lead to many non-cancer-specific deaths due to comorbidities, affecting overall survival (OS). Therefore, our study aims to identify independent prognostic factors for elderly PC patients using the Surveillance, Epidemiological, and End Results (SEER) database and develop and validate nomograms for specific survival (CSS) and OS in elderly PC patients, as well as to provide a reference basis for the clinical diagnosis and treatment work.

PATIENTS AND METHODS

Data Source and Data Extraction

We downloaded patient data from the SEER database, including patients aged 65 years and older diagnosed with PC between 2010 and 2018. The SEER database is a national cancer database containing 18 cancer registries covering approximately 30% of the population. Since the patient information in the SEER database is anonymized and the data is publicly available, ethical approval and patient informed consent were not required for our study. The research methodology used in this study follows the research guidelines published in the SEER database.

We collected clinicopathological information for all elderly PC patients, including age, race, year of diagnosis, marital status, histological tumor grade, TNM stage, surgery, radiotherapy, chemotherapy, PSA, and biopsy GS. Patient follow-up results,

including survival status, cause of death, and survival time, are also available from the SEER database. Inclusion criteria: (1) patients age≥ 65;(2) with a pathological diagnosis of PC. Exclusion criteria: (1) patients younger than 65 years old; (2) tumor grade is unknown; (3) TNM stage is unknown; (4) surgical method is unknown; (5) PSA is not clear; (6) survival time is less than one month or survival time is unknown. The flowchart of patient inclusion and exclusion is shown in **Figure 1**.

Patients were classified as white, Black, and other types (American Indian/AK Native, Asian/Pacific Islander). The histological classification of tumors includes high differentiation (grade I), moderate differentiation (grade II), low differentiation (grade III), and undifferentiated (grade IV). According to the SEER surgical code, the surgical methods are divided into non-surgical surgery (surgical code 0), local tumor resection (surgical code 10-30), and radical prostatectomy (surgical code 50-80).

Development and Validation of the Nomograms

We first randomly divided the patients into two groups for the development and internal validation of the nomogram. All data were randomly divided into the training set (70%) and the validation set (30%). In addition, the data of PC patients from 2016-to 2018 were externally validated in time. Univariate and multivariate Cox proportional regression models were used to identify independent risk factors affecting patients' outcomes in the training set. We constructed nomograms based on a multivariate Cox regression model for predicting CSS and OS at 3-,5-, and 8-year. The calibration curves for 1,000 bootstrap samples were used to verify the accuracy of the nomograms. The consistency index (c-index) and the area under the subject operating characteristic curve (AUC) were used to test the accuracy and discrimination of the models.

Clinical Application

We use the decision analysis curve (DCA) to evaluate the clinical value of the nomograms for predicting CSS and OS at 3-,5-, and 8-years compared with the TNM staging system and D'Amico risk stratification system. Furthermore, we also calculated the risk for each patient from the nomograms. All patients were divided into high-risk and low-risk groups based on the cutoff value of the subject operating characteristic curve (ROC). The production principle of the ROC curve is to set several different critical values for the variable, calculate the corresponding sensitivity (sensitivity) and specificity (specificity) at each critical value, and then take the sensitivity as the ordinate and 1-specificity as the abscissa to draw the curve. Our ROC curves are time-dependent and are time-dependent. We used time-based ROC curves to test the nonlinear relationship of CSS or OS with risk values in the dataset and sought the closest risk score to HR=1 on the ROC curve as the cutoff. The Log-rank test and Kaplan-Meier (K-M) curves examined the differences in survival between high-risk and low-risk patients. In addition, surgical differences among patients in different risk groups were also analyzed.

Statistical Analysis

Continuous variables(age) were tested for normal distribution and conform to the normal distribution, described by the mean ± standard deviation. Chi-square or non-parametric U tests were used for comparison between groups. Other categorical variables were described by frequency (%), and the groups were compared using the chi-square test. The Cox regression models analyzed patient prognostic factors, and the log-rank test and K-M curves analyzed the survival differences of patients. All statistical methods were performed using R software version 4.1.0 and SPSS26.0. The R packages including "DynNom", "RMS", "Survival", and "ggDCA" were used. A P value less than 0.05 was considered statistically significant.

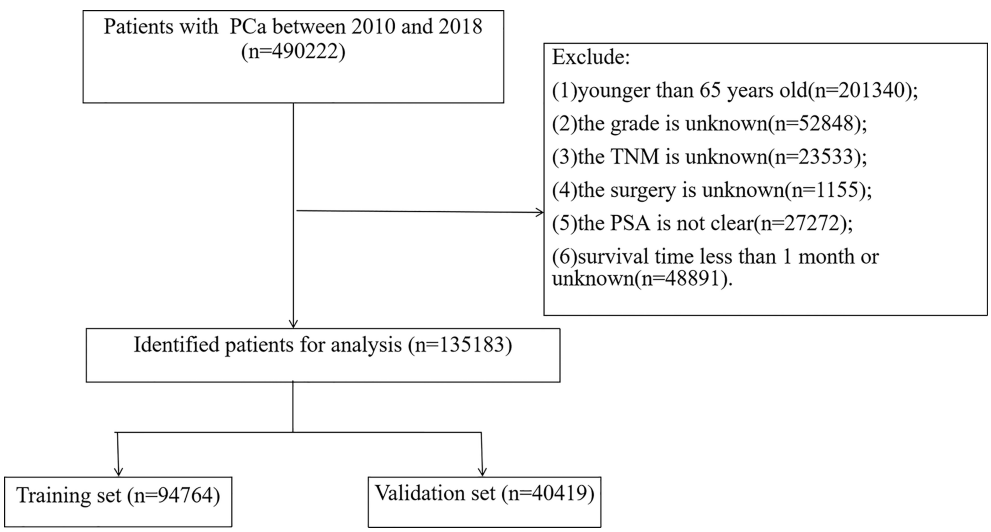


FIGURE 1 | Flowchart for inclusion and exclusion of elderly patients with PC.

RESULT

Clinical Features

A total of 135,183 patients between 2010 and 2018 were included in this study. All patients were randomly assigned to the training set (N=94764) and the validation set (N=40419). The mean age of both groups was 71.6 ± 5.1 years, and most of both groups were white (79.4%) and married (67.8%). The tumors included grade I (14.8%), II (40.3%), III (43.5%), and IV (1.44%).Patients with stage T1 (45.7%), T2 (39.1%), T3 (13.4%), and T4 (1.73%).Most patients were in stage N0 (95.8%) and staged M0 (95.5%).Patients with Non-surgical treatment (65.0%), patients who underwent local tumor resection (5.93%), and patients who underwent radical prostatectomy (29.1%).99% of patients received chemotherapy. 39.8% received radiotherapy, while 60.2% did not. Most patients had an unknown biopsy GS score (71.1%), and biopsy GS7 was about 20%. PSA 4-10ng/ml(57.1%), >10ng/ml(33.2%). The data did not show significant statistical bias in both groups, and the results are shown in **Table 1**.

Univariate and Multivariate COX Regression Analysis

Univariate Cox regression models were first used in training set to analyze and screen for factors associated with patient survival. The results showed that these factors, including age, race, marriage, tumor grade, TNM stage, surgery, chemotherapy, radiotherapy, PSA, and biopsy GS, could all affect patient survival. Then, multivariate Cox regression models were used to screen for independent risk factors associated with CSS and OS of elderly PC patients. The results showed that age, race, marriage, tumor grade, TNM stage, surgery, radiotherapy, chemotherapy, PSA, and biopsy GS were prognostic factors affecting patient CSS and OS. The analysis results are shown in **Tables 2** and **3**.

Nomograms Development for the 3-Year, 5-Year, and 8-Year CSS and OS

We constructed nomograms that predicted CSS and OS at 3-year,5-year, and 8-year in elderly PC patients based on multivariate Cox regression models **(Figure 2).** From the figure, age, TNM stage, tumor grade, surgery, PSA, and biopsy GS were the most influential factors for predicting CSS and OS in elderly PC patients. In addition, radiotherapy and chemotherapy are also essential factors. However, marriage and race had little effect on patient survival.

Validation of the Nomograms

Internal cross-validation was used to test the accuracy and discriminability of models. The C-index of the training set and the validation set for predicting CSS is 0.883(95%CI:0.877-0.889) and 0.887(95%CI:0.877-0.897), respectively. The C-index of the training set and the validation set for predicting OS was 0.77 (95%CI:0.766-0.774)and 0.767(95%CI:0.759-0.775), respectively. It indicated that the nomograms for CSS and OS have good recognition ability. In the training and validation set, the calibration curve shows that the predicted value of the nomograms for CSS and OS are highly consistent with the actual observed value **(Figure 3**). It shows that the nomograms have good accuracy. The AUC at 3-, 5-, and 8-years was 89.6,87.2, and 85.1, respectively, in the training set for CSS, and in the validation set for CSS, the AUC at 3-,5-, and 8-years was 89.9,88.4, and 85.7, respectively. In the training set for OS, the AUC at 3-, 5- and 8-years was 77.0,75.0, and 75.0, respectively, and in the validation set for OS, the AUC at 3-, 5- and 8-years was 77.4,75.4 and 74.5. The results show that the nomograms are very discriminative (**Figure 4**). The external validation set in time for predicting CSS was 0.903(95%CI:0.891-0.915), and the external validation set in time for predicting OS was 0.795(95%CI:0.785-0.805). The AUC at 1-and 2-year in the external validation set in time for CSS was 89.2 and 90.3, and the AUC in the external validation set in time for OS was 78.3 and 89.8 (**Figure S1**). Due to more than 75% of patients being classified as unknown GS groups, thus to bias the results, so we retrained the models after removing the unknown GS. The results showed that the C-index of the training set for CSS after removing the unknown GS is 0.785(95%CI:0.754-0.816), and the C-index of the training set for OS is 0.675(95%CI:0.657-0.693). The C-index of the validation set for CSS is 0.763(95%CI:0.708-0.818), and the C-index of the validation set for OS is 0.665(95% CI:0.640-0.690). Moreover, the AUC also showed that the model readiness and reliability would be decreased significantly after deleting the unknown GS (**Figure S2**).

Clinical Application of the Nomograms

In both the training set and the validation set for CSS and OS, DCA suggested that the nomograms had good clinical potential value (**Figure 5**). The nomograms for CSS at the 3,5,8-year validation set showed the best clinical potential value, followed by the D'Amico risk stratification and TNM staging systems. The nomogram for CSS at 3,5-year also showed the best clinical potential value in the training set. In contrast, the nomogram for CSS at 8-year had no apparent advantages over the other two, indicating that the nomogram is close to the other two models in the long term and does not show apparent advantages. The nomogram for OS at 3,5,8 years showed the best application potential in both the training and validation set, followed by D'Amico risk stratification and TNM staging. Based on the nomogram, we calculated each patient's risk value and the optimal cutoff value using the ROC curve. Patients were classified into the high-risk group (total score ≥293.59) and the low-risk group (total score <293.59) for predicting CSS, and patients were divided into the high-risk group (total score ≥184.88) and the low-risk group (total score <184.88) for predicting OS. The K-M curve showed that the CSS and OS rate of the patients in the high-risk group was significantly lower than that in the low-risk group both in the training and validation set (**Figure 6**). The 3-year, 5-year, and 8-year CSS rates of the patients in the high-risk group were 93.2%, 89.6%, and 84.7%, respectively. The low-risk group's 3-year, 5-year, and 8-year CSS rates were 99.6%, 99.2%, and 98.2%, respectively. The 3-year, 5-year, and 8-year OS rates of the patients in the high-risk group were 86.0%.0,76.7%, and 62.2%, respectively. The low-risk group's 3-year, 5-year, and 8-year OS rates were 96.8%, 93.9%, and 87.8%, respectively. We found that patients in the high-risk

TABLE 1 | Clinicopathological characteristics of elderly patients with PCa.

	All N = 135183	Training cohort N = 94764	Validation cohort N = 40419	p
Age	71.6 (5.51)	71.5 (5.48)	71.6 (5.53)	0.159
Race:				0.483
white	107381 (79.4%)	32025 (79.2%)	75356 (79.5%)	
black	16005 (11.8%)	4826 (11.9%)	11179 (11.8%)	
other	11797 (8.73%)	3568 (8.83%)	8229 (8.68%)	
Marital:				0.331
No	43594 (32.2%)	13111 (32.4%)	30482 (32.2%)	
Married	91590 (67.8%)	27308 (67.6%)	64282 (67.8%)	
Grade:				0.252
I	20021 (14.8%)	5880 (14.5%)	14141 (14.9%)	
II	54426 (40.3%)	16308 (40.3%)	38118 (40.2%)	
III	58791 (43.5%)	17666 (43.7%)	41125 (43.4%)	
IV	1945 (1.44%)	565 (1.40%)	1380 (1.46%)	
T:				0.516
T1	61819 (45.7%)	18482 (45.7%)	43337 (45.7%)	
T2	52878 (39.1%)	15889 (39.3%)	36989 (39.0%)	
T3	18142 (13.4%)	5368 (13.3%)	12774 (13.5%)	
T4	2344 (1.73%)	680 (1.68%)	1664 (1.76%)	
N:				0.603
N0	129460 (95.8%)	38726 (95.8%)	90734 (95.7%)	
N1	5723 (4.23%)	1693 (4.19%)	4030 (4.25%)	
M:				0.501
M0	129046 (95.5%)	38560 (95.4%)	90486 (95.5%)	
M1	6137 (4.54%)	1859 (4.60%)	4278 (4.51%)	
Surgery:				0.351
No	87843 (65.0%)	26354 (65.2%)	61489 (64.9%)	
Local tumor excision	8019 (5.93%)	2348 (5.81%)	5671 (5.98%)	
Radical prostatectomy	39321 (29.1%)	11717 (29.0%)	27604 (29.1%)	
Chemotherapy:				0.885
No	133822 (99.0%)	40015 (99.0%)	93807 (99.0%)	
Yes	1361 (1.01%)	404 (1.00%)	957 (1.01%)	
Radiation:				0.530
No	81407 (60.2%)	24288 (60.1%)	57119 (60.3%)	
Yes	53776 (39.8%)	16131 (39.9%)	37645 (39.7%)	
Gleason:				0.486
≤6	5141 (3.80%)	1549 (3.83%)	3592 (3.79%)	
3+4	17182 (12.7%)	5185 (12.8%)	11997 (12.7%)	
4+3	9542 (7.06%)	2819 (6.97%)	6723 (7.09%)	
≥8	7217 (5.34%)	2102 (5.20%)	5115 (5.40%)	
Unknown	96101 (71.1%)	28764 (71.2%)	67337 (71.1%)	
PSA:				0.245
<4	13169 (9.74%)	3966 (9.81%)	9203 (9.71%)	
4-10	77140 (57.1%)	23168 (57.3%)	53972 (57.0%)	
>10	44874 (33.2%)	13285 (32.9%)	31589 (33.3%)	
CSS:				0.680
Dead	6184 (4.57%)	1864 (4.61%)	4320 (4.56%)	
Alive	128999 (95.4%)	38555 (95.4%)	90444 (95.4%)	
Survival.months	46.2 (29.9)	46.3 (29.9)	46.1 (29.9)	0.146

group had the highest CSS and OS rate for undergoing radical prostatectomy, but most patients did not receive surgery. Most patients in the low-risk group underwent radical prostatectomy or non-surgical treatment, with no significant difference in patient CSS and OS rate (**Figure 7**).

DISCUSSION

This study developed a nomogram using a population-based SEER database to predict CSS in elderly PC patients. With 135,183 patients included in this study, we successfully developed a nomogram to predict 3,5, and 8-year CSS in elderly PC patients, while internal validation demonstrated good calibration and discrimination of this nomogram. The nomogram consists of multiple independent prognostic factors, including age, marriage, Race, PSA, biopsy GS, radiotherapy, chemotherapy, surgical, tumor grade, and TNM stage.

In the past 170 years, PC has evolved from a rare disease to the most common non-cutaneous cancer; with the popularity of PSA screening, more and more patients can be detected in the early stages of the disease. PSA is the most common tumor

TABLE 2 | Univariate and multivariate analyses of CSS in training cohort.

	Univariate			Multivariate		
	HR	**95%CI**	**P**	**HR**	**95%CI**	**P**
Age	1.12	1.12-1.13	<0.001	1.055	1.051-1.059	<0.001
Race						
white						
black	1.29	1.18-1.41	<0.001	1.174	1.089-1.264	<0.001
other	0.8	0.71-0.89	<0.001	0.662	0.599-0.732	<0.001
Marital						
No						
Married	0.66	0.63-0.71	<0.001	0.849	0.807-0.895	<0.001
Grade						
I						
II	1.87	1.48-2.36	<0.001	1.521	1.262-1.835	<0.001
III	9.78	7.85-12.18	<0.001	3.963	3.307-4.749	<0.001
IV	20.3	13.63-30.22	<0.001	6.414	4.58-8.983	<0.001
T						
T1						
T2	1.13	1.05-1.21	0.001	1.196	1.127-1.27	<0.001
T3	1.43	1.3-1.57	<0.001	1.389	1.271-1.518	<0.001
T4	11.37	10.3-12.55	<0.001	2.216	2.019-2.432	<0.001
N						
N0						
N1	8.13	7.54-8.77	<0.001	1.365	1.269-1.468	<0.001
M						
M0						
M1	27.62	25.96-29.4	<0.001	7.254	6.773-7.769	<0.001
Surgery						
No						
Local tumor excision	2.25	2.06-2.45	<0.001	1.609	1.494-1.732	<0.001
Radical prostatectomy	0.24	0.21-0.27	<0.001	0.759	0.541-1.064	0.109
Chemotherapy						
No						
Yes	10.64	9.42-12.02	<0.001	1.41	1.268-1.567	<0.001
Radiation						
No						
Yes	0.64	0.6-0.68	<0.001	0.647	0.609-0.688	<0.001
PSA						
<4						
4-10	0.71	0.62-0.81	<0.001	0.798	0.711-0.895	<0.001
>10	4.16	3.66-4.73	<0.001	1.559	1.396-1.741	<0.001
Gleason						
≤6						
3+4	0.88	0.55-1.41	0.592	0.511	0.348-0.748	0.001
4+3	1.35	0.83-2.2	0.231	0.608	0.405-0.913	0.016
≥8	8.18	5.4-12.4	<0.001	2.419	1.711-3.422	<0.001
Unknown	10.43	7.04-15.45	<0.001	2.699	1.712-4.253	<0.001

marker in PC screening, although high levels of PSA in benign prostatic hyperplasia and prostatitis reduce PSA specificity as a cancer marker. However, many studies show that high levels of PSA are associated with a poor prognosis in prostate cancer patients (31). However, a previous study reported that low levels of PSA decreased the CSS in PC patients (32). Interestingly, different predictive models have opposite conclusions on the effects of PSA for the bone metastasis patients in PC, and the Indonesian nomogram suggests that higher PSA levels are associated with a worse prognosis (33). The SEER database also confirmed that high PSA levels were associated with poor prognosis (34). However, the Japanese nomogram suggests that PC patients with higher PSA suggest a better prognosis (35).In previous studies, PSA was generally divided into three grades: <4,4-10 and> 10 ng/ml (36, 37). This standard also graded our study. We also confirmed that PSA> 10 ng/ml was associated with a poor prognosis, consistent with most previous reports. However, the patient prognosis of PSA< 4 ng/ml and PSA 4-10ng/ml was inconsistent with the expression level of PSA. Considering the particularity of elderly PC patients, the PSA of most patients may be high, and the patients with PSA< 4 ng/ml are very few, which may bias the results. Secondly, it is reported that PSA 4-10 ng/ml itself is a "gray area" in PC (38), so the previous PSA classification criteria for elderly PC patients can not accurately respond to the actual situation. Moreover, we also made the corresponding nomogram for predicting OS, which showed the same conclusion that patients with PSA< 4 ng/ml had worse OS than PSA 4-10 ng/ml. Considering that it is

TABLE 3 | Univariate and multivariate analyses of OS in training cohort.

	Univariate			Multivariate		
	HR	**95%CI**	**P**	**HR**	**95%CI**	**P**
Age	1.12	1.11-1.12	<0.001	1.073	1.07-1.075	<0.001
Race						
white						
black	1.29	1.22-1.35	<0.001	1.245	1.192-1.301	<0.001
other	0.69	0.65-0.75	<0.001	0.632	0.595-0.672	<0.001
Marital						
No						
Married	0.66	0.64-0.68	<0.001	0.798	0.774-0.822	<0.001
Grade						
I						
II	1.24	1.14-1.36	<0.001	1.158	1.078-1.244	<0.001
III	2.57	2.37-2.79	<0.001	1.767	1.645-1.898	<0.001
IV	3.78	2.84-5.03	<0.001	2.487	1.966-3.146	<0.001
T						
T1						
T2	0.96	0.92-1	0.04	1.127	1.089-1.166	<0.001
T3	0.86	0.81-0.92	<0.001	1.217	1.147-1.292	<0.001
T4	4.41	4.07-4.77	<0.001	1.889	1.754-2.034	<0.001
N						
N0						
N1	3.51	3.31-3.74	<0.001	1.261	1.188-1.338	<0.001
M						
M0						
M1	8.82	8.42-9.24	<0.001	3.241	3.087-3.403	<0.001
Surgery						
No						
Local tumor excision	1.89	1.79-1.99	<0.001	1.373	1.308-1.44	<0.001
Radical prostatectomy	0.33	0.31-0.35	<0.001	0.635	0.519-0.776	<0.001
Chemotherapy						
No						
Yes	4.7	4.21-5.23	<0.001	1.393	1.269-1.529	<0.001
Radiation						
No						
Yes	0.85	0.82-0.88	<0.001	0.716	0.692-0.741	<0.001
PSA						
<4						
4-10	0.8	0.75-0.85	<0.001	0.884	0.837-0.935	<0.001
>10	2.25	2.11-2.4	<0.001	1.283	1.213-1.358	<0.001
Gleason						
≤6						
3+4	0.7	0.6-0.81	<0.001	0.582	0.512-0.662	<0.001
4+3	0.72	0.61-0.86	<0.001	0.597	0.515-0.691	<0.001
≥8	1.58	1.35-1.84	<0.001	0.918	0.799-1.055	0.228
Unknown	2.91	2.58-3.29	<0.001	1.155	0.93-1.435	0.192

inconsistent with clinical practice, we also recommend subsequent researchers adopt a new PSA classification standard for elderly PC patients over 65 years of age: PSA <10 ng/ml,10-20 ng/ml,> 20 ng/ml, instead of 4 ng/ml as the intermediate value, which may avoid outcome bias. SEER database-based studies have divided PSA into three levels: <10 ng/ml,10-20 ng/ml, and> 20 ng/ml. Most prediction models confirmed that PSA greater than 10 rather than 4 is associated with a poor prognosis (28, 34).

GS as an essential tool for predicting the prognosis of patients with PC has been revised multiple times since being proposed. The most common risk stratification for prostate cancer is the D'Amico classification, also used by the National Comprehensive Cancer Network (39), which divides Gleason scores into three Gleason score groups (2 – 6,7 and 8 – 10). However, the current GS system still has vast defects, especially with a total score of 7. The patient prognosis of GS 3 + 4 and GS 4 + 3 is very different, so the simple GS does not accurately predict the prognosis of PC patients. Therefore, a study developed a nomogram for patients with GS 4 + 3 and GS 3 + 4, which showed that patients with GS 4 + 3 had a worse prognosis than patients with GS 3 + 4 (28), and our study reached the same conclusion.Meanwhile, our nomogram showed that GS> 8 is associated with a worse prognosis, consistent with previous reports, but patients with GS<6 have a worse prognosis than those with GS7. We consider that the majority of patients did not undergo a needle biopsy for PC, resulting in more than 75% of cases being classified as biopsy GS unknown group, thus bias the results. After removing the

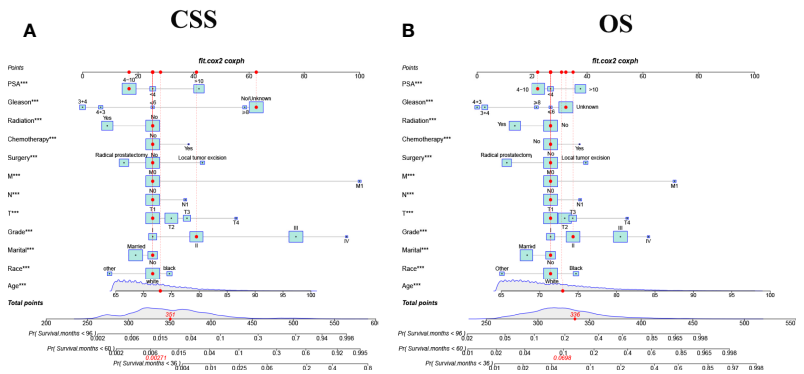


FIGURE 2 | The nomograms for predicting 3-,5-,8-year CSS and OS in elderly patients with PC. **(A)** The nomogram for predicting CSS in elderly patients with PC. **(B)** The nomogram for predicting OS in elderly patients with PC. ***P < 0.001 vs. reference.

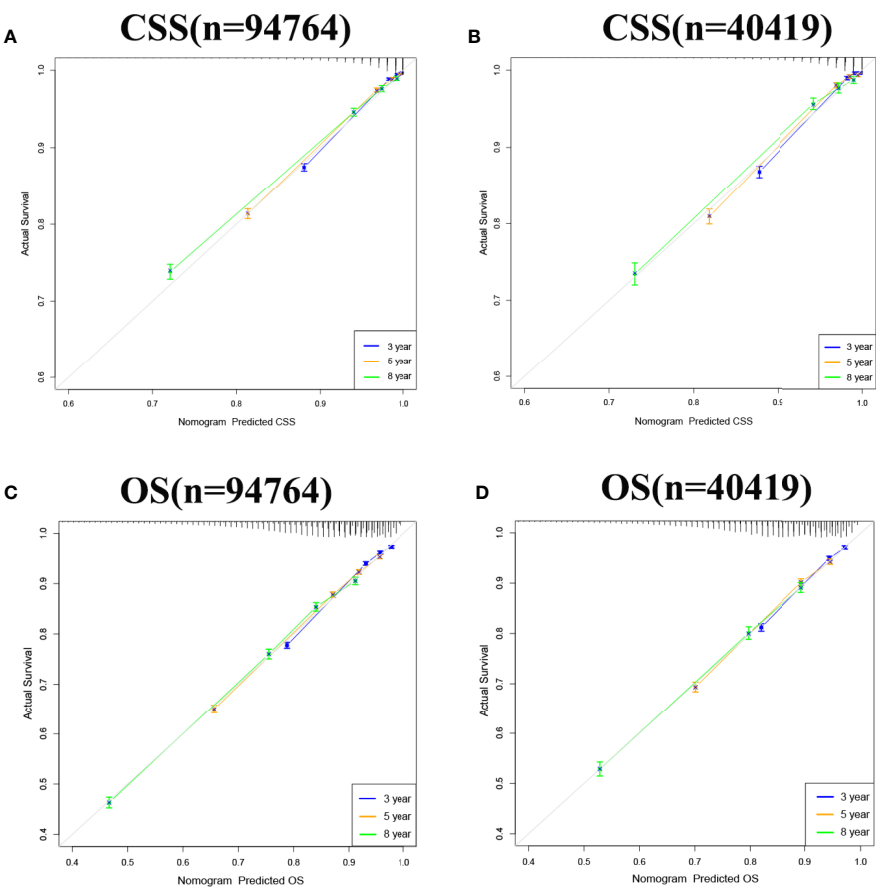


FIGURE 3 | Calibration curve of the nomograms for predicting 3-,5-,8-year CSS and OS in elderly patients with PC. **(A)** Calibration curve of the nomograms for predicting 3-,5-,8-year CSS in the training set. **(B)** Calibration curve of the nomograms for predicting 3-,5-, and 8-year CSS in the validation set. **(C)** Calibration curve of the nomograms for predicting 3-,5-,8-year OS in the training set. **(D)** Calibration curve of the nomograms for predicting 3-,5-,8-year OS in the validation set. The horizontal axis is the predicted value in the nomogram, and the vertical axis is the observed value.

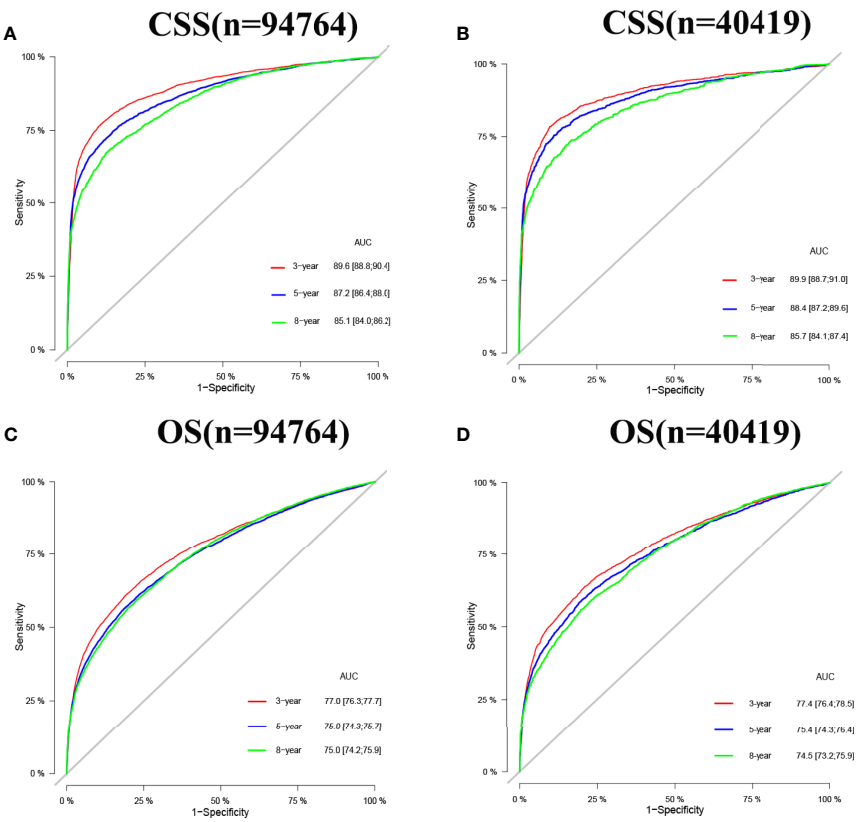


FIGURE 4 | AUC for predicting 3-, 5-, and 8-year CSS and OS in elderly patients with PC. **(A)** The AUC at 3-, 5-, and 8-year for CSS in the training set was 89.6,87.2, and 85.1. **(B)** The AUC for CSS in the validation set was 89.9,88.4, and 85.7. **(C)** The AUC at 3-, 5- and 8-year for OS in the training set was 77.0,75.0, and 75.0. **(D)** The AUC at 3-, 5- and 8-year for OS in the validation set was 77.4,75.4, and 74.5.

unknown GS, the model was trained again. The result showed that deleting this part of the patients caused a significant decrease in the accuracy and reliability of the model and, therefore, poor availability.

PC treatment uses active monitoring, surgical resection, and androgen deprivation (ADT), combined with radiotherapy and chemotherapy. The surgical methods of PC mainly include: radical prostatectomy (RP) and local tumor resection (LTR), our prediction model showed that patients with RP had better outcomes than patients treated without surgery, which is consistent with previous reports (40, 41), and patients with LTR had the worst prognosis, considering many low-risk patients only need active monitoring rather than surgical treatment can obtain good prognosis, and many elderly PC patients who need RP but choose palliative surgery due to inability of tolerating prolonged general anesthesia, life expectancy is less than 10 years, etc. Therefore, our study found that for elderly patients with PC, patients with local tumor resection instead had a worse prognosis than those who did not receive surgical treatment. Radiation therapy (RT) is a conventional treatment method for PC patients, and it is mainly used clinically to treat patients with medium-risk or high-risk nonmetastatic prostate cancer (42, 43). RT is noninvasive and does not require considering the cardiorespiratory risks arising from systemic or local anesthesia. Therefore, it can be used to treat intolerable elderly patients with PC. At the same time, it does not require hospitalization, improving patient compliance while also reducing hospitalization costs. More importantly, RT can avoid some side effects of surgery, such as urinary incontinence. Our findings showed that elderly PC patients with RT had better outcomes than patients without RT, consistent with previous studies (44). Thus, RT is becoming a key component of multimodal therapy at multiple stages of PC. Unlike most solid tumors, chemotherapy (CT) is not the primary treatment for PC. Almost all PC will eventually develop metastatic castration-resistant prostate cancer (mCRPC), and in the 1970s and 1980s, although many chemotherapeutic drugs were tested in CRPC. Most drugs were tested in phase II clinical trials, and although many seemed promising, none were ultimately shown to prolong survival (45). As most of the patients did not receive chemotherapy, our predictive model also did not show a survival advantage of chemotherapy for elderly PC patients.

Cancer is now commonly evaluated through the tumor-lymph node-metastasis (TNM) system, which was previously considered the "gold standard" for staging and a benchmark for

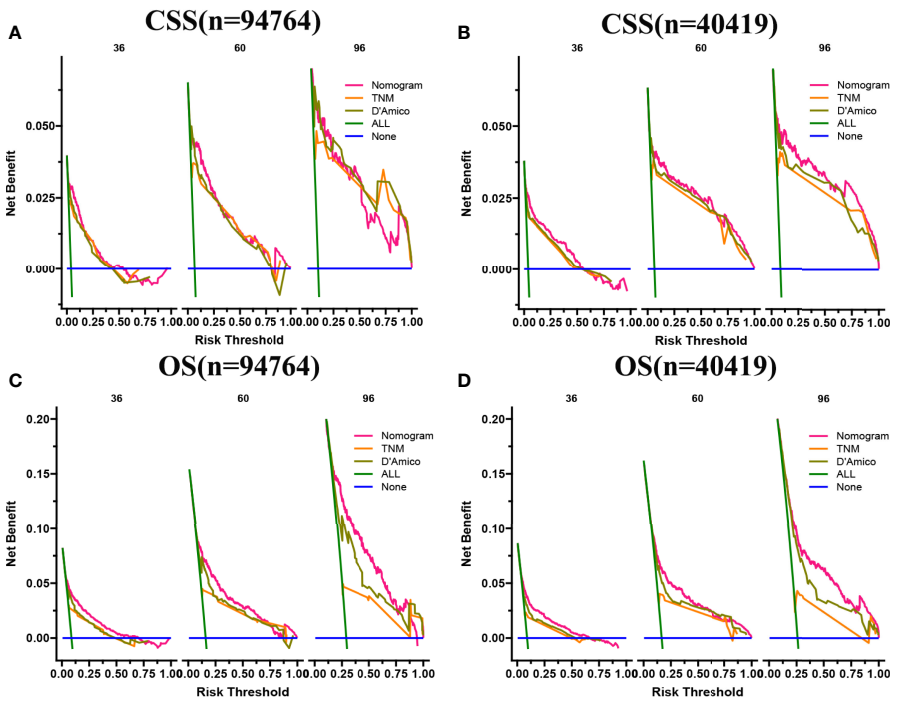


FIGURE 5 | DCA of the nomograms for predicting CSS and OS. **(A)** The nomogram for CSS at 3,5-year showed the best clinical potential value in the training set, while the nomogram for CSS at 8-year had no apparent advantages over the other two. **(B)** In the validation set, the nomogram for CSS at the 3,5,8-year showed the best clinical potential value, followed by the D'Amico risk stratification system and TNM staging system. **(C, D)** The nomogram for OS at 3-,5-,8-year showed the best application potential in both the training and validation sets, followed by D'Amico risk stratification and TNM staging.

prognosis (46). Most nomograms of PC prognosis (28, 34) included traditional TNM stages, showing that T4 has the worst prognosis compared to other T stages, and patients with distant and lymph node metastasis had a worse prognosis compared to patients without distant and lymph node metastasis, which is consistent with our findings.

The social support provided by marital relationships can promote a healthy lifestyle and increase healthcare-seeking behavior, so marriage is associated with favorable outcomes for most cancer patients. Charlotte Salmon et al. emphasized the elevated risk of PC in single men (47). Libby Ellis et al. also demonstrated that marital status in prostate cancer patients is associated with prognosis (48). Our prediction model also shows the relationship between marital status and prognosis, proving that married patients often indicate a good prognosis, which may be related to the influence of marriage on mood, a medical decisions, etc. In addition, it may be due to the responsibility of married families. Married cancer patients can detect physical abnormalities and actively cooperate with treatment (49).In terms of race or ethnicity, previous studies showed that black men had the highest PC incidence and mortality rates (50). David A Siegel et al. reported 5-year survival (6) between 2001 and 2016, showing that 5-year survival was higher among other ethnic men (84.4%) and white men (82.8%) than among black men (79.1%). Our results showed that elderly Black and white patients with PC had a worse prognosis than other races, consistent with previous reports. Although survival rates of PC patients of different ages vary by stage, however, compared with younger patients, elderly patients likely secondary to the rapid development of resistant PC, reduced ability to receive available treatment, and the effects of comorbidities often have lower long-term survival (51), which is also supported by our results.

Although the nomogram based on the SEER database has good accuracy, there are potential limitations. First is the lack of critical clinicopathological variables, such as smoking, alcohol consumption, hemoglobin, etc. In addition, for PC patients, PSA is an important indicator related to prognosis. However, it was not included in the SEER database until 2010, so we can only choose the data after 2010 for building the prediction model. Meanwhile, ADT, as one of the non-surgical treatment options for PC patients, is usually used for high-risk local or systemic advanced disease that is not suitable for radical surgery. However, it lacks ADT-related data in the SEER database, so our model also lacks the relationship between ADT and prognosis.

Furthermore, database-based studies are all retrospective, which may risk selection bias. Further multi-center prospective studies with a large sample are needed to validate this nomogram. Finally, although our nomogram did not consider all related prognostic variables, we still included key variables and validated them, so there would not be a significant deviation.

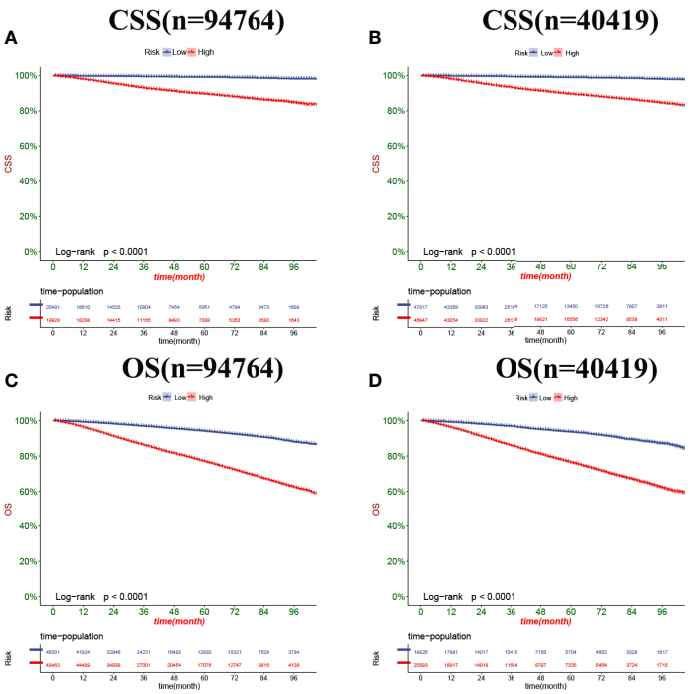


FIGURE 6 | Kaplan-Meier curves of patients in the low-risk and high-risk groups. The K-M curve showed that the CSS rate of the patients in the high-risk group was significantly lower than that in the low-risk group both in the training set **(A)** and validation set **(B)**. The K-M curve showed that the OS rate of the patients in the high-risk group was significantly lower than that in the low-risk group both in the training set **(C)** and validation set **(D)**.

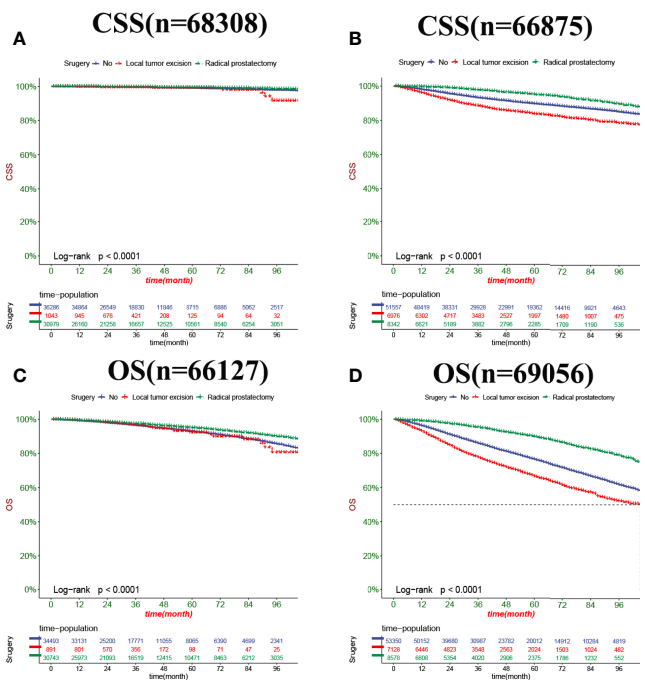


FIGURE 7 | Kaplan-Meier curves of patients with a different surgery. **(A)** The CSS rate of patients in the low-risk group underwent different surgery. **(B)** The CSS rate of patients in the high-risk group underwent different surgery. **(C)** The OS rate of patients in the low-risk group underwent different surgery. **(D)** The OS rate of patients in the high-risk group underwent different surgery.

CONCLUSION

We explored the factors influencing CSS and OS in elderly patients with PC and found that age, race, marriage, PSA, biopsy GS, surgical approach, radiotherapy, chemotherapy, tumor grade, and TNM stage were independent risk factors affecting patients' CSS and OS. We established nomograms to predict the CSS and OS in elderly patients with PC. The models have been internally validated with good accuracy and reliability, and they can help make better clinical decisions for clinicians and patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZZ and CZ designed the study. ZZ, JW, JL, MJL, LJ and MXL collected and analyzed the data. ZZ drafted the initial manuscript. CZ, TM, JL, XT and DH revised the article critically. XW, XT, CZ, JL, DH, MJL reviewed and edited the article. All authors approved the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Special Key Project of Chongqing Technology Innovation and Application Development (No. Cstc2019jscx-tjsbX0003), Yunnan Education Department of Science Research Fund (No. 2020 J0228), Kunming City Health Science and Technology Talent "1000" training Project (No. 2020- SW (Reserve)-112), Kunming Health and Health Commission Health Research Project (No. 2020-0201-001), and Kunming Medical Joint Project of Yunnan Science and Technology Department (No. 202001 AY070001-271).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.918780/full#supplementary-material

Supplementary Figure 1 | The AUC at 1-,2-year in the external validation set in time. A: The AUC at 1-,2-year in the external validation set for CSS was 89.2 and 90.3. B: The AUC in the external validation set for OS was 78.3 and 89.8.

Supplementary Figure 2 | The AUC at 3-,5-,8-year for predicting CSS and OS after deleting the unknown GS. A: The AUC at 3-,5-,8-year for predicting CSS after deleting the unknown GS was 76.4,82.6, and 82.3.B: The AUC at 3-,5-,8-year for predicting OS after deleting the unknown GS was 65.9,77.9 and 73.8.

REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2022. *CA Cancer J Clin* (2022) 72(1):7–33. doi: 10.3322/caac.21708
2. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer Treatment and Survivorship Statistics, 2019. *CA Cancer J Clin* (2019) 69(5):363–85. doi: 10.3322/caac.21565
3. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to Build a Bridge From a Population-Based to a More "Personalized" Approach to Cancer Staging. *CA Cancer J Clin* (2017) 67(2):93–9. doi: 10.3322/caac.21388
4. Srigley JR, Delahunt B, Samaratunga H, Billis A, Cheng L, Clouston D, et al. Controversial Issues in Gleason and International Society of Urological Pathology (ISUP) Prostate Cancer Grading: Proposed Recommendations for International Implementation. *Pathology* (2019) 51(5):463–73. doi: 10.1016/j.pathol.2019.05.001
5. Liu Y, Xia Q, Xia J, Zhu H, Jiang H, Chen X, et al. The Impact of Marriage on the Overall Survival of Prostate Cancer Patients: A Surveillance, Epidemiology, and End Results (SEER) Analysis. *Can Urol Assoc J* (2019) 13(5):E135–9. doi: 10.5489/cuaj.5413
6. Siegel DA, O'Neil ME, Richards TB, Dowling NF, Weir HK. Prostate Cancer Incidence and Survival, by Stage and Race/Ethnicity - United States, 2001-2017. *MMWR Morb Mortal Wkly Rep* (2020) 69(41):1473–80. doi: 10.15585/mmwr.mm6941a1
7. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur Urol* (2016) 69(3):428–35. doi: 10.1016/j.eururo.2015.06.046
8. Gleason DF, Mellinger GT. Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *J Urol* (1974) 111(1):58–64. doi: 10.1016/S0022-5347(17)59889-4
9. Bjartell A. Words of Wisdom. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Eur Urol* (2006) 49(4):758–9. doi: 10.1016/j.eururo.2006.02.007
10. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am J Surg Pathol* (2016) 40(2):244–52. doi: 10.1097/PAS.0000000000000530
11. Kundu SD, Roehl KA, Yu X, Antenor JA, Suarez BK, Catalona WJ. Prostate Specific Antigen Density Correlates With Features of Prostate Cancer Aggressiveness. *J Urol* (2007) 177(2):505–9. doi: 10.1016/j.juro.2006.09.039
12. Albertsen PC. Prostate Cancer Screening and Treatment: Where Have We Come From and Where are We Going? *BJU Int* (2020) 126(2):218–24. doi: 10.1111/bju.15153
13. Mahal BA, Yang DD, Wang NQ, Alshalalfa M, Davicioni E, Choeurng V, et al. Clinical and Genomic Characterization of Low-Prostate-Specific Antigen, High-Grade Prostate Cancer. *Eur Urol* (2018) 74(2):146–54. doi: 10.1016/j.eururo.2018.01.043
14. Force U.S.P.S.T., Grossman DC, Curry SJ, Owens DK, Bibbins-Domingo K, Caughey AB, et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* (2018) 319(18):1901–13. doi: 10.1001/jama.2018.3710
15. Wilt TJ, Dahm P. PSA Screening for Prostate Cancer: Why Saying No is a High-Value Health Care Choice. *J Natl Compr Canc Netw* (2015) 13 (12):1566–74. doi: 10.6004/jnccn.2015.0182
16. Gnanapragasam VJ, Lophatananon A, Wright KA, Muir KR, Gavin A, Greenberg DC. Improving Clinical Risk Stratification at Diagnosis in Primary Prostate Cancer: A Prognostic Modelling Study. *PLos Med* (2016) 13(8):e1002063. doi: 10.1371/journal.pmed.1002063
17. Zelic R, Garmo H, Zugna D, Stattin P, Richiardi L, Akre O, et al. Predicting Prostate Cancer Death With Different Pretreatment Risk Stratification Tools: A Head-To-Head Comparison in a Nationwide Cohort Study. *Eur Urol* (2020) 77(2):180–8. doi: 10.1016/j.eururo.2019.09.027
18. Iasonos A, Schrag D, Raj GV, Panageas KS. How to Build and Interpret a Nomogram for Cancer Prognosis. *J Clin Oncol* (2008) 26(8):1364–70. doi: 10.1200/JCO.2007.12.9791
19. Sternberg CN. Are Nomograms Better Than Currently Available Stage Groupings for Bladder Cancer? *J Clin Oncol* (2006) 24(24):3819–20. doi: 10.1200/JCO.2006.07.1290
20. Gittleman H, Sloan AE, Barnholtz-Sloan JS. An Independently Validated Survival Nomogram for Lower-Grade Glioma. *Neuro Oncol* (2020) 22(5):665–74. doi: 10.1093/neuonc/noz191

21. Wu S, Zheng J, Li Y, Yu H, Shi S, Xie W, et al. A Radiomics Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. *Clin Cancer Res* (2017) 23(22):6904–11. doi: 10.1158/1078-0432.CCR-17-1510
22. Ljungberg B. Kidney Cancer: A New Nomogram Predicting Survival in Renal Cell Carcinoma. *Nat Rev Urol* (2010) 7(8):423–4. doi: 10.1038/nrurol.2010.117
23. Kim SY, Cho N, Choi Y, Lee SH, Ha SM, Kim ES, et al. Factors Affecting Pathologic Complete Response Following Neoadjuvant Chemotherapy in Breast Cancer: Development and Validation of a Predictive Nomogram. *Radiology* (2021) 299(2):290–300. doi: 10.1148/radiol.2021203871
24. Loupakis F, Moretto R, Aprile G, Muntoni M, Cremolini C, Iacono D, et al. Clinico-Pathological Nomogram for Predicting BRAF Mutational Status of Metastatic Colorectal Cancer. *Br J Cancer* (2016) 114(1):30–6. doi: 10.1038/bjc.2015.399
25. Gafita A, Calais J, Grogan TR, Hadaschik B, Wang H, Weber M, et al. Nomograms to Predict Outcomes After (177)Lu-PSMA Therapy in Men With Metastatic Castration-Resistant Prostate Cancer: An International, Multicentre, Retrospective Study. *Lancet Oncol* (2021) 22(8):1115–25. doi: 10.1016/S1470-2045(21)00274-6
26. Zhao J, Sun G, Liao B, Zhang X, Armstrong CM, Yin X, et al. Novel Nomograms for Castration-Resistant Prostate Cancer and Survival Outcome in Patients With De Novo Bone Metastatic Prostate Cancer. *BJU Int* (2018) 122(6):994–1002. doi: 10.1111/bju.14398
27. Lee C, Light A, Alaa A, Thurtle D, van der Schaar M, Gnanapragasam VJ. Application of a Novel Machine Learning Framework for Predicting non-Metastatic Prostate Cancer-Specific Mortality in Men Using the Surveillance, Epidemiology, and End Results (SEER) Database. *Lancet Digit Health* (2021) 3 (3):e158–65. doi: 10.1016/S2589-7500(20)30314-9
28. Zhu X, Gou X, Zhou M. Nomograms Predict Survival Advantages of Gleason Score 3+4 Over 4+3 for Prostate Cancer: A SEER-Based Study. *Front Oncol* (2019) 9:646. doi: 10.3389/fonc.2019.00646
29. Droz JP, Albrand G, Gillessen S, Hughes S, Mottet N, Oudard S, et al. Management of Prostate Cancer in Elderly Patients: Recommendations of a Task Force of the International Society of Geriatric Oncology. *Eur Urol* (2017) 72(4):521–31. doi: 10.1016/j.eururo.2016.12.025
30. Sharma S. Imaging and Intervention in Prostate Cancer: Current Perspectives and Future Trends. *Indian J Radiol Imaging* (2014) 24(2):139–48. doi: 10.4103/0971-3026.134399
31. Ang M, Rajcic B, Foreman D, Moretti K, O'Callaghan ME. Men Presenting With Prostate-Specific Antigen (PSA) Values of Over 100 Ng/mL. *BJU Int* (2016) 117 Suppl 4:68–75. doi: 10.1111/bju.13411
32. Song P, Yang B, Peng Z, Zhou J, Ren Z, Fang K, et al. Reduced Cancer-Specific Survival of Low Prostate-Specific Antigen in High-Grade Prostate Cancer: A Population-Based Retrospective Cohort Study. *Int J Surg* (2020) 76:64–8. doi: 10.1016/j.ijsu.2020.02.024
33. Halabi S, Small EJ, Kantoff PW, Kattan MW, Kaplan EB, Dawson NA, et al. Prognostic Model for Predicting Survival in Men With Hormone-Refractory Metastatic Prostate Cancer. *J Clin Oncol* (2003) 21(7):1232–7. doi: 10.1200/JCO.2003.06.100
34. Liu D, Kuai Y, Zhu R, Zhou C, Tao Y, Han W, et al. Prognosis of Prostate Cancer and Bone Metastasis Pattern of Patients: A SEER-Based Study and a Local Hospital Based Study From China. *Sci Rep* (2020) 10(1):9104. doi: 10.1038/s41598-020-64073-6
35. Armstrong AJ, Garrett-Mayer E, de Wit R, Tannock I, Eisenberger M. Prediction of Survival Following First-Line Chemotherapy in Men With Castration-Resistant Metastatic Prostate Cancer. *Clin Cancer Res* (2010) 16 (1):203–11. doi: 10.1158/1078-0432.CCR-09-2514
36. Schymura MJ, Sun L, Percy-Laurry A. Prostate Cancer Collaborative Stage Data Items–Their Definitions, Quality, Usage, and Clinical Implications: A Review of SEER Data for 2004-2010. *Cancer* (2014) 120 Suppl 23:3758–70. doi: 10.1002/cncr.29052
37. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, et al. Prevalence of Prostate Cancer Among Men With a Prostate-Specific Antigen Level < or =4.0 Ng Per Milliliter. *N Engl J Med* (2004) 350(22):2239–46. doi: 10.1056/NEJMoa031918
38. Tang P, Du W, Xie K, Deng X, Fu J, Chen H, et al. Transition Zone PSA Density Improves the Prostate Cancer Detection Rate Both in PSA 4.0-10.0 and 10.1-20.0 Ng/Ml in Chinese Men. *Urol Oncol* (2013) 31(6):744–8. doi: 10.1016/j.urolonc.2011.06.012
39. D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, Broderick GA, et al. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. *JAMA* (1998) 280(11):969–74. doi: 10.1001/jama.280.11.969
40. Bill-Axelson A, Holmberg L, Garmo H, Taari K, Busch C, Nordling S, et al. Radical Prostatectomy or Watchful Waiting in Prostate Cancer - 29-Year Follow-Up. *N Engl J Med* (2018) 379(24):2319–29. doi: 10.1056/NEJMoa1807801
41. Hamdy FC, Donovan JL, Lane JA, Mason M, Metcalfe C, Holding P, et al. 10-Year Outcomes After Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *N Engl J Med* (2016) 375(15):1415–24. doi: 10.1056/NEJMoa1606220
42. Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment With Curative Intent. *Eur Urol* (2021) 79(2):243–62. doi: 10.1016/j.eururo.2020.09.042
43. Cornford P, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer. Part II-2020 Update: Treatment of Relapsing and Metastatic Prostate Cancer. *Eur Urol* (2021) 79(2):263–82. doi: 10.1016/j.eururo.2020.09.046
44. Philippou Y, Sjoberg H, Lamb AD, Camilleri P, Bryant RJ. Harnessing the Potential of Multimodal Radiotherapy in Prostate Cancer. *Nat Rev Urol* (2020) 17(6):321–38. doi: 10.1038/s41585-020-0310-3
45. Hurwitz M. Chemotherapy in Prostate Cancer. *Curr Oncol Rep* (2015) 17 (10):44. doi: 10.1007/s11912-015-0468-7
46. Varma M, Cochlin D, Delahunt B, Kynaston H, Rees J, Rous B, et al. TNM Clinical Staging of Prostate Cancer: Issues and Solutions. *BJU Int* (2019) 123 (3):382–4. doi: 10.1111/bju.14589
47. Salmon C, Song L, Muir K, Collaborators U, Pashayan N, Dunning AM, et al. Marital Status and Prostate Cancer Incidence: A Pooled Analysis of 12 Case-Control Studies From the PRACTICAL Consortium. *Eur J Epidemiol* (2021) 36(9):913–25. doi: 10.1007/s10654-021-00781-1
48. Ellis L, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL. Racial and Ethnic Disparities in Cancer Survival: The Contribution of Tumor, Sociodemographic, Institutional, and Neighborhood Characteristics. *J Clin Oncol* (2018) 36(1):25–33. doi: 10.1200/JCO.2017.74.2049
49. Enewold L, Harlan LC, Tucker T, McKenzie S. Pancreatic Cancer in the USA: Persistence of Undertreatment and Poor Outcome. *J Gastrointest Cancer* (2015) 46(1):9–20. doi: 10.1007/s12029-014-9668-x
50. Reddy S, Shapiro M, Morton RJr., Brawley OW. Prostate Cancer in Black and White Americans. *Cancer Metastasis Rev* (2003) 22(1):83–6. doi: 10.1023/A:1022216119066
51. Bernard B, Burnett C, Sweeney CJ, Rider JR, Sridhar SS. Impact of Age at Diagnosis of De Novo Metastatic Prostate Cancer on Survival. *Cancer* (2020) 126(5):986–93. doi: 10.1002/cncr.32630

Conflict of Interest: 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.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 04 July 2022
doi: 10.3389/fonc.2022.876967

Auto-Segmentation Ultrasound-Based Radiomics Technology to Stratify Patient With Diabetic Kidney Disease: A Multi-Center Retrospective Study

Jifan Chen[12.†], Peile Jin[1,2†], Yue Song[1,2†], Liting Feng[3†], Jiayue Lu[4], Hongjian Chen[1,2,5], Lei Xin[1,2], Fuqiang Qiu[1,2], Zhang Cong[1,2], Jiaxin Shen[1,2], Yanan Zhao[1,2], Wen Xu[1,2], Chenxi Cai[6], Yan Zhou[7], Jinfeng Yang[6], Chao Zhang[1,2], Qin Chen[3], Xiang Jing[7]* and Pintong Huang[1,2,8]**

[1] Department of Ultrasound in Medicine, The Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China, [2] Ultrasound in Medicine and Biomedical Engineering Research Center, The Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China, [3] Department of Ultrasound, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China, [4] Department of Clinical Laboratory, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [5] Post-Doctoral Research Center, Hangzhou Supor South Ocean Pharmaceutical Co., Ltd, Hangzhou, China, [6] Department of Ultrasound, The People's Hospital of Yinshang, Anhui, China, [7] Department of Ultrasound, Tianjin Third Central Hospital, Tianjin, China, [8] Research Center for Life Science and Human Health, Binjiang Institute of Zhejiang University, Hangzhou, China

Background: An increasing proportion of patients with diabetic kidney disease (DKD) has been observed among incident hemodialysis patients in large cities, which is consistent with the continuous growth of diabetes in the past 20 years.

Purpose: In this multicenter retrospective study, we developed a deep learning (DL)-based automatic segmentation and radiomics technology to stratify patients with DKD and evaluate the possibility of clinical application across centers.

Materials and Methods: The research participants were enrolled retrospectively and separated into three parts: training, validation, and independent test datasets for further analysis. DeepLabV3+ network, PyRadiomics package, and least absolute shrinkage and selection operator were used for segmentation, extraction of radiomics variables, and regression, respectively.

Results: A total of 499 patients from three centers were enrolled in this study including 246 patients with type II diabetes mellitus (T2DM) and 253 patients with DKD. The mean intersection-over-union (Miou) and mean pixel accuracy (mPA) of automatic segmentation of the data from the three medical centers were 0.812 ± 0.003, 0.781 ± 0.009, 0.805 ± 0.020 and 0.890 ± 0.004, 0.870 ± 0.002, 0.893 ± 0.007, respectively. The variables from the renal parenchyma and sinus provided different information for the diagnosis and follow-up of DKD. The area under the curve (AUC) of the radiomics model for differentiating between DKD and T2DM patients was 0.674 ± 0.074 and for differentiating between the high and low stages of DKD was 0.803 ± 0.037.

OPEN ACCESS

Edited by:
Jian Lu,
Peking University Third Hospital, China

Reviewed by:
Zhenyu Liu,
Institute of Automation (CAS), China
Mohammad Hasan,
Indiana University, United States
Yangsean Choi,
The Catholic University of Korea, South Korea

***Correspondence:**
Pintong Huang
huangpintong@zju.edu.cn
Xiang Jing
dr.jingxiang@aliyun.com
Qin Chen
1718686103@qq.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

Received: 16 February 2022
Accepted: 31 May 2022
Published: 04 July 2022

Citation:
Chen J, Jin P, Song Y, Feng L, Lu J, Chen H, Xin L, Qiu F, Cong Z, Shen J, Zhao Y, Xu W, Cai C, Zhou Y, Yang J, Zhang C, Chen Q, Jing X and Huang P (2022) Auto-Segmentation Ultrasound-Based Radiomics Technology to Stratify Patient With Diabetic Kidney Disease: A Multi-Center Retrospective Study. Front. Oncol. 12:876967. doi: 10.3389/fonc.2022.876967

Conclusion: In this study, we developed a DL-based automatic segmentation, radiomics technology to stratify patients with DKD. The DL technology was proposed to achieve fast and accurate anatomical-level segmentation in the kidney, and an ultrasound-based radiomics model can achieve high diagnostic performance in the diagnosis and follow-up of patients with DKD.

Keywords: ultrasound, radiomics, deep learning, diabetic kidney disease, multicenter

INTRODUCTION

Diabetic kidney disease (DKD) is a common microvascular complication in patients with diabetes and is the primary cause of kidney failure in ∼40% of diabetic patients (1, 2). In China, an increasing proportion of patients with DKD has been observed among incident hemodialysis patients in large cities, which is consistent with the continuous growth in diabetic patients in the past 20 years (3). DKD diagnosis is based on estimated glomerular filtration rate (eGFR), urinary abnormalities such as proteinuria and microhematuria, and kidney biopsy, which is often avoided in the early stages of DKD.

In patients with suspected kidney function injury, ultrasound imaging is the first imaging technique to be performed for the diagnosis and follow-up of its progression (4). Researchers have demonstrated that certain parameters such as cortical echogenicity and thickness in B-mode ultrasonography, resistance index in color Doppler sonography (4, 5), elastography scores (6) and time-intensity curve parameters in contrast-enhanced ultrasound imaging (7) can effectively reflect the kidney function in patients with chronic kidney disease (CKD). Ultrasound is frequently applied as an available and noninvasive technology for the diagnosis and follow-up of DKD in patients suffering from type II diabetes mellitus (T2DM) for a long duration. However, a conventional ultrasound examination is limited owing to the visual grayscale image, which reduces its potential for identifying a large amount of valuable information. Furthermore, the interpretation of ultrasound images is variable and unreliable owing to inexperienced sonographers, especially in the diffused form of the disease.

Radiomics is a rapidly growing discipline based on quantitative image analysis that reflects image textures and morphology using gray values, which provides a quantitative, solid, and objective foundation for analytic standardization to inform clinical decisions (8). In radiomics technology, a delineated region of interest (ROI) is vital for extracting radiomics variables; however, the accurate anatomical segmentation of ROIs is a time-consuming and experience-dependent process. Recent advances in image segmentation, classification, and registration through deep learning (DL) have considerably expanded the scope and scale of medical image analysis (9). A state-of-the-art network "DeepLabV3+" was reported for semantic image segmentation, which achieved high accuracy when compared to other networks (10).

Therefore, this multicenter retrospective study aimed to achieve an automatic anatomical-level segmentation of the kidney in T2DM patients with/without DKD and to build an ultrasound-image-based radiomics model for diagnosis and follow-up of patients with different stages of DKD. This method can extensively utilize the information contained in conventional ultrasound images and achieve acceptable accuracy, resulting in a quick process that uses easily available resources and demonstrates the potential for further clinical use.

MATERIALS AND METHODS

Study Design and Patients

This study was registered at ClinicalTrials.gov No. NCT05025540 and the informed consent requirement was waived due to the retrospective study design. This multicenter retrospective study was approved by the ethics consultant committee of the Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU), the People's Hospital of Yingshang (PHYS) and Tianjin Third Central Hospital (TJTCH), and 162 DKD and 131 T2DM patients were consecutively enrolled in this study between January 2016 and December 2020 and January 2018 and December 2020, respectively, as the control group from SAHZU. Moreover, 35 DKD and 52 T2DM patients from the People's Hospital of Yingshang (PHYS) and 56 DKD and 63 T2DM patients from Tianjin Third Central Hospital (TJTCH) were also enrolled consecutively.

The following parameters were used to define DKD: 1) urinary albumin-to-creatinine ratio (UACR) > 30 mg/24 h, with an increase greater than twice the original value in three subsequent examinations conducted over 3-6 months; 2) eGFR

Abbreviations: DKD, diabetic kidney disease; DL, deep learning; DN, diabetic nephropathy; T2DM, type II diabetes mellitus; Miou, mean intersection-over-union; mPA, mean pixel accuracy; AUC, area under the curve; ROI, region of interest, CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate, SAHZU, Second Affiliated Hospital of Zhejiang University School of Medicine; PHYS, People's Hospital of Yingshang; TJTCH: Tianjin Third Central Hospital; UACR, urinary albumin-to-creatinine ratio; UAER, urinary albumin excretion rate; LASSO, Least absolute shrinkage and selection operator; ICC, intraclass correlation coefficient; CUDA, Compute Unified Device Architecture; BMI, body mass index; HbA1c, glycated hemoglobin A1c; FBG, fasting blood-glucose; ACR, Albumin-to-Creatinine Ratio; GLCM, gray-level co- occurrence matrix; GLRLM, gray-level run-length matrix; GLSZM, gray-level size zone matrix; GLDZM, gray-level distance zone matrix; NGTDM, neighborhood gray tone difference matrix; NGDLM, neighboring gray-level dependence matrix; ROC, receiver operating characteristic; WHFE, wavelet-H_Firstorder_Entropy; GGL, gradient_glcm_lmc2; WHGS, wavelet-H_glrlm_shortRunEmphasis; GSHGLE, glrlm_shortRunHighGrayLevelEmphasis; GGLNUN, glszm_GrayLevelNonUniformyNormalized; WHGGLV, wavelet-H_gldm_GrayLevelVariance.

< 60 ml min^{-1} for more than three months; 3) pathological result of kidney biopsy shows evidence of DKD.

Clinical Stage of Diabetic Kidney Disease (DKD)

The clinical stage of DKD was defined based on the Chinese guidelines for the diagnosis and treatment of DKD (11, 12). DKD stage I, called as high-filtration stage, was defined as having normal or a marginally elevated eGFR (> 90 mL/min/1.73 m^2) and negative microalbuminuria; DKD stage II, called as microalbumin stage, was defined as having a urinary albumin excretion rate (UAER) of approximately 20-200 μg/min or 30-300 mg/24 h and eGFR > 60 mL/min/1.73 m^2; DKD stage III, called as massive albuminuria stage, was defined as having normal UACR >300 mg/g, UAER > 200 ug/min or >300 mg/24 h and eGFR > 15 mL/min/1.73 m^2; DKD stage IV, called as renal failure stage, was defined as having eGFR < 15 mL/min/1.73 m^2. In this study, we defined the low DKD stage as the stage lower than stage II and high DKD stage as the stage higher than stage III.

The low-stage DKD was defined as the DKD stage that is lower than stage III; moreover, high-stage DKD was defined as the DKD stage that is higher than or equal to stage III.

Kidney Ultrasound Scan

A 3-5 MHz convex probe was used for the adult kidney scans. For the kidney, patients with left/right lateral or dorsal positions were scanned in the coronal plane, and B-mode images were recorded. The maximum longitudinal kidney images with the renal sinus and parenchyma were saved. Both sides of the kidney were recorded per patient.

Inclusive and Exclusive Criteria

Inclusive Criteria

Patients were enrolled according to the following criteria: 1) patients with a clinical diagnosis of T2DM and DKD; 2) patients with clear B-mode ultrasound images on both sides of the kidney (left and right); 3) there were no missing values of the selected clinical data such as eGFR and UACR in the electronic medical records.

Exclusive Criteria

The following criteria were considered for excluding patients: 1) patients with large kidney-space-occupying diseases, such as kidney renal cysts and tumors; 2) ultrasound images with severe shadows or incomplete kidney borders.

Data Extraction and Model Building

First, we manually and automatically delineated the ROI of the renal parenchyma and sinus.

The radiologists for ROI are at least three years in ultrasound diagnosis. Besides, the labelme software in python is used to draw ROIs. Then, collected images was transformed into JPG image format and imported into PyCharm software. Third, the radiomics data were extracted using the Python package PyRadiomics (13) and 1682 radiomics variables from each side of the kidney (841 parenchyma and 841 sinus) were extracted in this study. Moreover, we calculated the intraclass correlation coefficient (ICC) between the extracted data from the manually and automatically delineated ROIs. The ICC values of variables that were higher than 0.7 were selected for further model building. Least absolute shrinkage and selection operator (LASSO) regression was used to select the significant features (14). Finally, the radiomics scores were calculated and the diagnostic performance was compared using the receiver operating characteristic (ROC) curve (**Figure 1A**).

Deep Learning (DL) Algorithm

DeepLabV3+ improves pyramid-shaped hole pooling, cascaded multiple hole convolutions, and extensively uses batch normalization. First, DeepLabV3+ uses an atrous spatial pyramid pooling structure to mine multiscale contextual content information. The decoding structure gradually reconstructs the spatial information to capture object boundaries more effectively. Second, a new decoding module was added to reconstruct the boundary information. Third, we attempted to use MobileNet module as the backbone of the network to reduce the number of parameters and increase the speed of the network. The network structure of DeepLabV3+ is shown in **Figure 1B**.

Data Enhancement

In the training set for DeepLabV3+, we first used the data enhancement strategies to extensively use ultrasound images, such as random horizontal flip, random scale change, and random Gaussian blur. After data enhancement, a five-fold increase in the number of pictures was achieved.

Experimental Environment

The DeepLabV3+ network was built using PyTorch version 1.9.0 with Compute Unified Device Architecture (CUDA) version 11.1 (15). NVIDIA GeForce RTX3070 Ti platform was used in a Windows 10 operating system. Statistical modeling (LASSO) was performed using R and RStudio.

Statistical Analysis

Continuous data with normal distribution are shown as mean ± standard deviation, and data with a non-normal distribution are shown as median (quartile interval). Categorical data are expressed as a number (percentage). The distributions of our data were measured using the Shapiro-Wilk test. In univariate analysis, continuous data were compared using Student's t-test, one-way analysis of variance, Mann–Whitney U test, or Kruskal–Wallis H test, and categorical variables were compared using the χ^2 test. Multiple comparisons were performed using Tukey correction.

RESULTS

Baseline Characteristics Among the Three Medical Centers

A total of 499 patients were enrolled in this study: 131 T2DM patients and 53, 91, 18 patients with DKD stages II, III, IV, respectively, from SAHZU; 63 T2DM patients and 25, 31 patients with DKD stages II, III, respectively, from TJTCH; 52

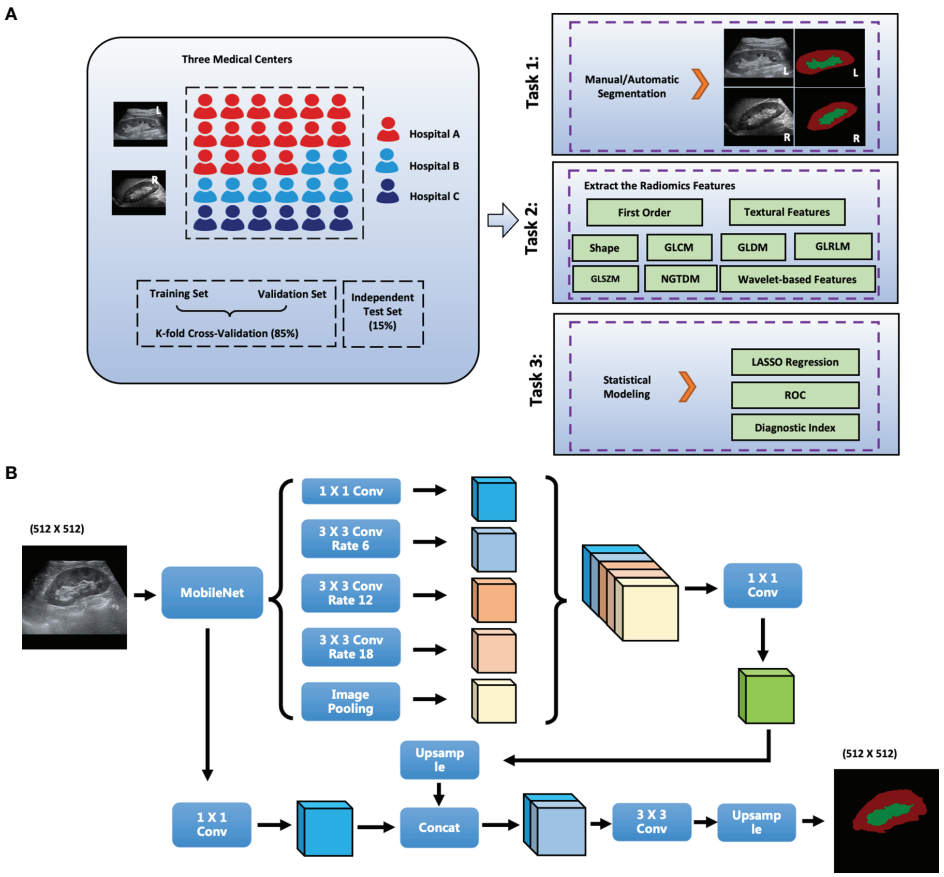


FIGURE 1 | The scheme of this study. **(A)** Flowchart of the Study; **(B)** Network Structure of DeepLabV3+; L, left; R, right; GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; First_order, first order statistics; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix. Conv: Convolution layer; Hospital A: The Second Affiliated Hospital of Zhejiang University School of Medicine, SAHZU; Hospital B: Tianjin Third Central Hospital, THTCH; Hospital C: The People's Hospital of Yingshang, PHYS.

T2DM patients and 13, 20, 2 patients with DKD stages II, III, IV, respectively, from PHYS.

The average ages were 57 (51–64), 59 (48-65), 57 (46-69) years for T2DM patients and 60 (51-68), 63 (51-75), 60 (49-70) years for DKD patients among the three datasets, respectively. Moreover, the percentages of male patients were 61.8, 57.1, and 25.0% among the T2DM patients and 59.3, 51.8, and 42.9% among the DKD patients in the three datasets, respectively.

In addition, among the three datasets, the average fasting blood glucose levels were 8.8 (6.4-12.2), 7.46 (5.72-9.99), 9.7 (7.0-12.0) mmol/L in T2DM patients, 7.4 (5.6-10.1), 8.8 (5.6-12.1), 9.4 (7.4-11.9) mmol/L in DKD patients, respectively. Further, among the three datasets, the eGFRs were 105.8 (94.8-115.8), 104.7 (85.7-123.8),100.2 (84.2-107.8) ml/min/1.73m^2 in T2DM patients, 69.0 (33.3-105.3), 83.9 (47.0-99.1), 83.4 (47.9-100.4) ml/min/1.73m^2 in DKD patients, respectively (**Table 1**).

DL-Based Anatomical-Level Segmentation

Two radiologists independently delineated the kidney border, renal parenchyma, and renal sinus in ultrasound images, and the inconsistency was resolved through discussions. The DeepLabV3+ network was applied as an automatic anatomical-level segmentation technology, whose structure is illustrated in **Figure 1B**. As shown in **Figure 2**, the trained DeepLabV3+ model showed good segmentation ability in patients with clear ultrasound images (Patient Nos. 1 and 2). To further verify the robustness and accuracy of DL technology, we tested the model using the ultrasound images of patients with inferior ultrasound images (Patient Nos. 3 and 4). The trained model showed that it could compensate for the missing border caused by inferior ultrasound windows and maintain high segmentation accuracy.

Moreover, we verified the segmentation ability of DeepLabV3+ on a separate test set (N=50) from SAHZU and two independent test datasets from TJTCH and PHYS. From **Figure 3**, **Table 2**, it can be observed that the mean intersection-over-union (Miou) and mean pixel accuracy (mPA) of the SAHZU test set were 0.812 ± 0.003 and 0.890 ± 0.004, respectively. Moreover, the Miou and mPA of the TJTCH dataset were 0.781 ± 0.009 and 0.870 ± 0.002, respectively, and those of the PHYS dataset were 0.805 ± 0.020 and 0.893 ± 0.007, respectively (**Table 2**). These results demonstrate

TABLE 1 | The Basic Characteristics of Study Patients in Three Medical Centers. SAHZU.

Variables	SAHZU		TJTCH		PHYS		*P*
T2DM	131 (44.7%)		63 (52.9%)		52 (59.8%)		< 0.05
DKD Stage II	53 (18.1%)		25 (21.0%)		13 (14.9%)		
DKD Stage III	91 (31.1%)		31 (26.1%)		20 (23.0%)		
DKD Stage IV	18 (6.1%)		0 (0.0%)		2 (2.3%)		
Demographics:	**T2DM**	**DKD**	**T2DM**	**DKD**	**T2DM**	**DKD**	***P***
Age	57 (51-64)	60 (51-68)	59 (48-65)	64 (53-75)	57 (46-69)	61 (50-72)	$P_1^{\#}$ P_2*
Male (%)	81 (61.8%)	96 (59.3%)	36 (57.1%)	29 (51.8%)	13 (25.0%)	15 (42.9%)	P_1* $P_2^{\#}$
BMI	24.5 (22.6-26.1)	25.3 (23.4-27.2)	25.9 (22.4-29.3)	26.1 (22.8-29.4)	25.6 (22.2-29.0)	25.7 (22.2-29.2)	P_1* $P_2^{\#}$
Hypertension	56 (42.7%)	130 (80.2%)	36 (57.1%)	41 (73.2%)	20 (38.5%)	18 (51.4%)	$P_1^{\#}$ P_2*
DM duration	3285 (1825-5475)	3650 (2555-7118)	2555 (1095-4745)	3650 (1095-6388)	2920 (1095-3650)	3650 (1060-5475)	$P_1^{\#}$ $P_2^{\#}$
Biochemical value	**T2DM**	**DKD**	**T2DM**	**DKD**	**T2DM**	**DKD**	
HbA1c	8.9 (7.8-10.3)	8.4 (7.0-10.1)	8.3 (7.3-10.5)	9.1 (7.0-11.2)	9.7 (7.6-11.8)	8.8 (6.7-10.9)	P_1* $P_2^{\#}$
FBG	8.8 (6.4-12.2)	7.4 (5.6-10.1)	7.5 (5.7-10.0)	8.6 (5.8-11.4)	9.7 (7.0-12.0)	9.3 (7.4-10.5)	P_1* $P_2^{\#}$
Urea nitrogen	4.3 (5.1-6.2)	8.2 (5.2-12.2)	4.7 (4.0-5.4)	6.5 (4.7-8.5)	5.7 (4.4-7.1)	7.2 (4.9-10.5)	P_1* P_2*
Creatinine	59.5 (47-71)	94 (60.5-200)	62 (55.5-70.5)	76 (59-113)	59.2 (47.9-70.6)	75 (61-110)	$P_1^{\#}$ $P_2^{\#}$
Uric acid	300 (257-350)	382 (307-449)	281 (229-331)	319 (200-437)	241 (161-322)	288 (236-364)	P_1* P_2*
ACR	13.5 (8.9-18.0)	429.0 (58.0-2653)	5.6 (3.8-13.1)	244.9 (72.1-1166.4)	10.6 (4.3-17.0)	411.1 (57.0-2070.3)	P_1* $P_2^{\#}$
eGFR	105.8 (94.8-115.8)	69.0 (33.3-105.3)	104.7 (85.7-123.8)	83.9 (47.0-99.1)	100.2 (84.2-116.2)	83.4 (47.9-100.4)	P_1* $P_2^{\#}$

**: P-value ≤ 0.05; #: P-value > 0.05; P_1: P-value of three datasets in T2DM groups; P_2: P-value of three datasets in DKD groups; SAHZU, The Second Affiliated Hospital of Zhejiang University School of Medicine; TJTCH, Tianjin Third Central Hospital; PHYS, The People's Hospital of Yingshang; T2DM, type 2 diabetes mellitus; DKD, Diabetic Kidney Disease; BMI, body mass index; DM, diabetes mellitus; HbA1c, glycated hemoglobin A1c; FBG, fasting blood-glucose; ACR, Albumin-to-Creatinine Ratio; eGFR, estimated glomerular filtration rat.*

that high robustness and accuracy can be achieved by DL-based technology, thus providing a faster method for delineating the ROI.

Correlation Between Extracted Radiomics Variables

In total, 3364 radiomics variables were extracted per patient, including 841 from the renal parenchyma and 841 from the renal sinus per kidney.

First, we calculated the correlation between the extracted variables from the parenchyma and sinus. The extracted variables were completely irrelevant, with ICC values of 0.236 (0.142-0.356) and 0.249 (0.133-0.374) in T2DM and DKD patients, respectively (**Figures 3A, B**). These results demonstrate that different parts of the kidney can provide different information.

Next, we calculated the ICC between radiomics variables extracted using manual and DL-based automatic methods to select robust variables for further analysis. As expected, the variables extracted using manual and automatic methods were highly correlated. Moreover, the median and interquartile range of ICC in the parenchyma and sinus were 0.871 (0.728-0.937) and 0.860 (0.779-0.927), respectively (**Figures 3C, D**). The radiomics variables with ICC > 0.7 were selected in this study; thus, 2066 radiomics variables, including 974 and 1092 variables extracted from parenchyma and sinus, respectively, were used in the model building step.

Utilization of Radiomics Variables From Parenchyma and Sinus for Stratifying DKD Patients

The kidneys of patients with a high DKD stage tend to show higher echogenicity in both the parenchyma and sinus (**Figure 4A**), which can provide evidence that the radiomics variables have the potential to stratify DKD patients.

To reduce the dimensions of the variables, the LASSO procedure was performed. A total of 94 variables were selected, including 24 from the feature class of gray level co-occurrence matrix, 12 from gray level run length matrix, 12 from gray level size zone matrix, 16 from first order statistics, 3 from neighboring gray tone difference matrix, 6 from gray level dependence matrix, and 21 from wavelet decompositions (**Figure 4B**, **Table 3**).

Note that 73 variables from parenchyma and 21 from sinus were selected (**Figure 4B**), as illustrated in the density plot (**Figure 4C**), and the value distribution of wavelet-H_Firstorder_Entropy (WHFE) and gradient_glcm_lmc2 (GGL) from the parenchyma were shifted to the left in the high DKD stage, whereas the value of wavelet-H_glrlm_shortRunEmphasis (WHGS) was shifted to the

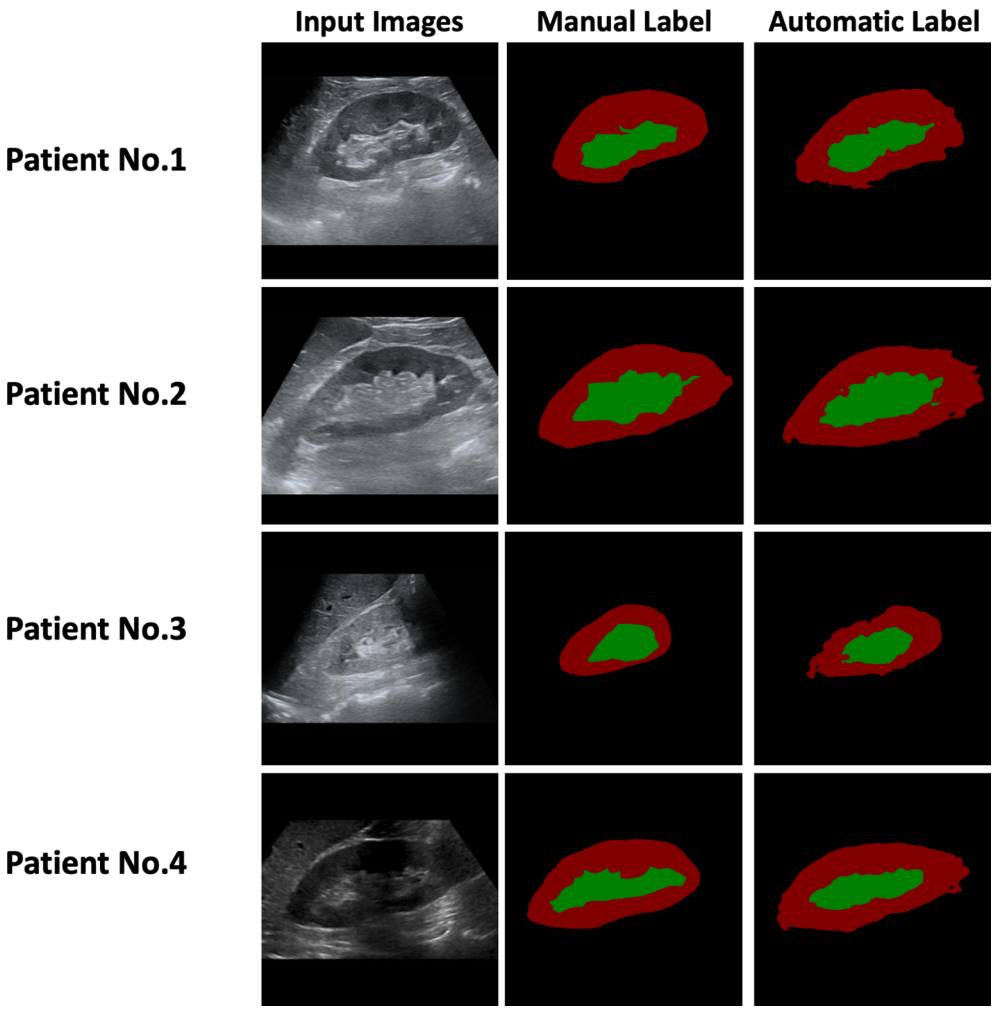


FIGURE 2 | Manual and Automatic Segmentation using Ultrasound Images of the Patients.

right in the high DKD stage. Moreover, the value of glrlm_shortRunHighGrayLevelEmphasis (GSHGLE) was shifted to the left in the high DKD stage, whereas the value of glszm_GrayLevelNonUniformyNormalized (GGLNUN) was shifted to the right in the high DKD stage. Further, the distribution of the value of wavelet-H_gldm_GrayLevelVariance (WHGGLV) in the high DKD stage was more spiculate than in the low DKD stage.

These results demonstrated that the variables extracted from both the parenchyma and sinus could provide positive diagnostic value in the diagnosis and follow-up of DKD, particularly in the low- and high-stage DKD (**Figure 4**).

Diagnostic Performance of Ultrasound-Based Radiomics in Stratifying DKD Patients

We randomly separated the data of the 499 patients from the three medical centers into three parts: 424 (85%) patients were divided into training and validation datasets and 75 (15%) patients were divided into independent test datasets. In the model building step, a k-fold cross-validation method was applied to the training and validation datasets to calculate the area under the curve (AUC), and the differences between the groups were compared. After the previous step, the best model was tested using independent test datasets, and the ROC and AUC were plotted and calculated.

While differentiating between DKD and T2DM patients, the radiomics model achieved moderate diagnostic performance with AUCs of 0.674 ± 0.074 in the parenchyma + sinus model, 0.6561 ± 0.0537 in the parenchyma model, and 0.6457 ± 0.0514 in the sinus model (**Figure 5A**). No statistical differences were found among the three models (all $P > 0.05$). In the independent test set, the AUCs of parenchyma + sinus, parenchyma, and sinus were 0.6779, 0.6536, and 0.6593, respectively (**Figure 5B**).

When differentiating between the high (≥ stage III) and low (≤ stage II) DKD stages, the radiomics model that combined the

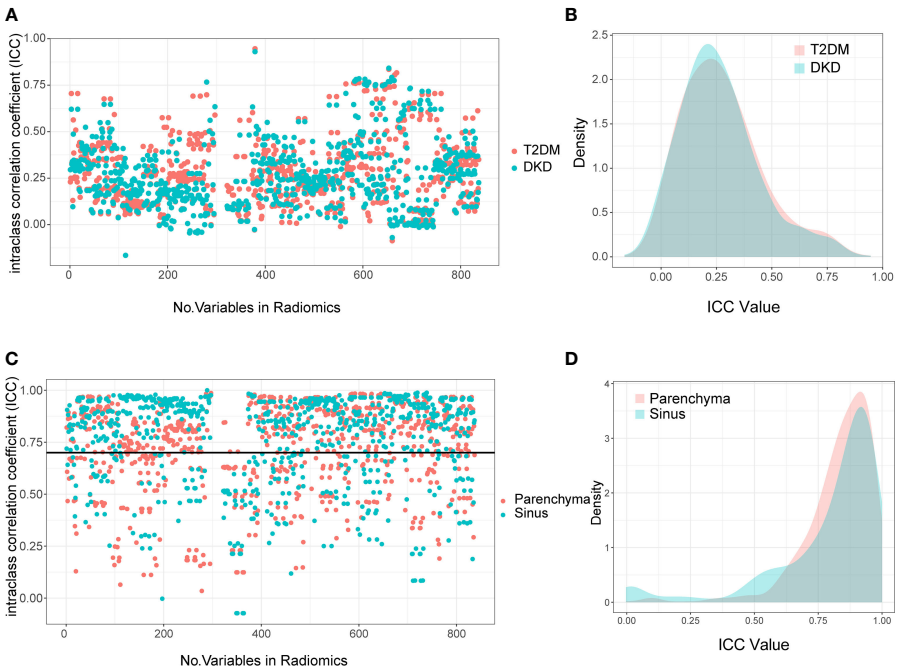


FIGURE 3 | Interclass Correlation Coefficients and Density Plots between Extracted Radiomics Variables. The Interclass Correlation Coefficients plot **(A)** and Density Plots **(B)** between parenchyma and sinus in T2DM and DKD group. The Interclass Correlation Coefficients plot **(C)** and Density Plots **(D)** between manual and automatic ROI drawing methods in T2DM and DKD group.

information from parenchyma and sinus achieved the highest diagnostic performance with AUC = 0.803 ± 0.037 (all $P<0.05$) after k-fold validation. Moreover, the AUCs of the radiomics model using only parenchyma and sinus variables were 0.75695 ± 0.038 and 0.716 ± 0.026, respectively (**Figure 5C**). In the independent test set, the AUCs of the models using parenchyma + sinus, parenchyma, and sinus variables were 0.8235, 0.7851, and 0.7304, respectively (**Figure 5D**).

In this study, the T2DM patients were a mix of T2DM patients without kidney function disorders and DKD stage I patients. These results demonstrated that the ultrasound images of patients in the early stage of DKD are similar to those of T2DM patients, which results in a moderate diagnostic performance of the radiomics model. However, the ultrasound-based radiomics model demonstrates good potential in differentiating between the low and high DKD stages, which is more useful for the stratification of patients with DKD.

TABLE 2 | Mean Intersection-over-union and Mean Pixel Accuracy in Three Medical Centers.

Dataset	Miou	mPA
SAHZU	0.812 ± 0.003	0.890 ± 0.004
TJTCH	0.781 ± 0.009	0.870 ± 0.002
PHYS	0.805 ± 0.020	0.893 ± 0.007

Miou, mean intersection-over-union; mPA, mean pixel accuracy; SAHZU, Second Affiliated Hospital of Zhejiang University School of Medicine; TJTCH, Tianjin Third Central Hospital; PHYS, People's Hospital of Yingshang.

Diagnostic Performance of DL-Based Automatic Segmentation, Radiomics for DKD

To further study the performance of the DL-based automatic segmentation, radiomics model in identifying DKD patients, we calculated and compared the AUC between the manual and automatic methods.

After k-fold cross-validation, the AUCs of the manual and automatic methods when differentiating between DKD and T2DM patients were 0.6797 ± 0.058 and 0.6626 ± 0.0547, respectively. Moreover, the AUCs of the manual and automatic methods of differentiating between patients at high and low DKD stages were 0.7967 ± 0.054 and 0.7732 ± 0.05478, respectively. There was no statistically significant difference between the AUCs of the manual and automatic methods (all P <0.05) Further, the AUCs of T2DM/DKD and high/low DKD stage while using manual and automatic methods were 0.692, 0.689, 0.8235, 0.7859, respectively, on the independent test set (**Figure 6**). The results demonstrated that manual and automatic segmentation, radiomics models achieved similar diagnostic performance.

DISCUSSION

Ultrasonography is an ideal evaluation tool and is widely used for the identification and analysis of several diseases (16). Although,

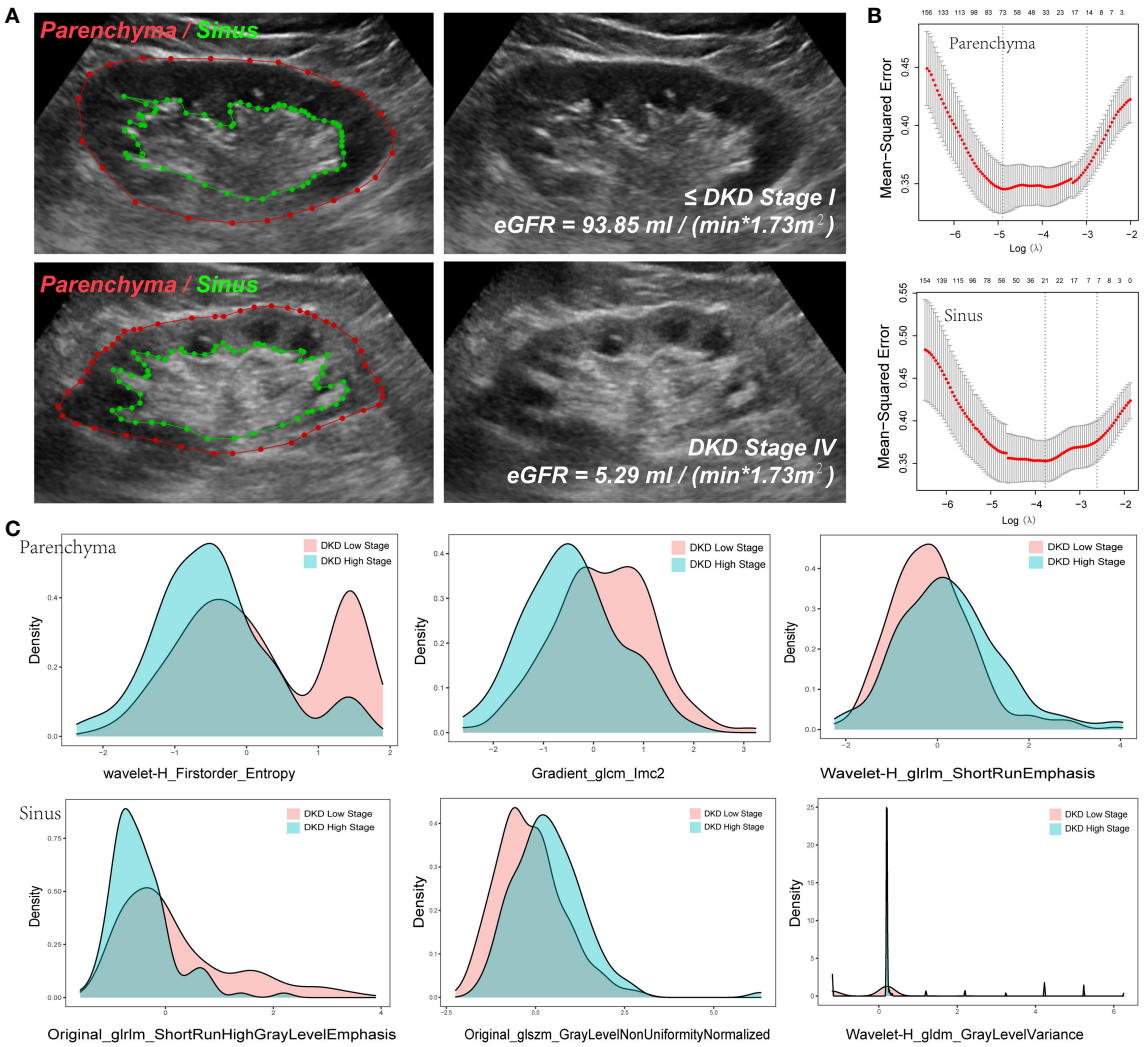


FIGURE 4 | Variables extracted from Parenchyma and Sinus. **(A)** The ROI of parenchyma and sinus in two DKD stage patients. **(B)** The Mean-square error plot of LASSO regression in parenchyma and sinus model. **(C)** The Density Plots between Extracted Radiomics Variables in parenchyma and sinus.

researchers can identify abnormal renal echogenicity, renal size, and other features in the diagnosis of kidney dysfunction through ultrasonography (17–19), the interpretation of ultrasound images by the naked eye is subjective and certain high-dimensional information can be missed; these problems can be solved using radiomics technology. However, accurate anatomical segmentation of the ROI, which is a time-consuming and experience-dependent process, is vital for effective utilization

TABLE 3 | Class of Extracted variables.

Variables Class	Sinus	Parenchyma
GLCM	7	17
GLRLM	2	10
GLSZM	3	9
First_order	5	11
NGTDM	1	2
GLDM	0	6
Wavelet	3	18

GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; First_order, first order statistics; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix.

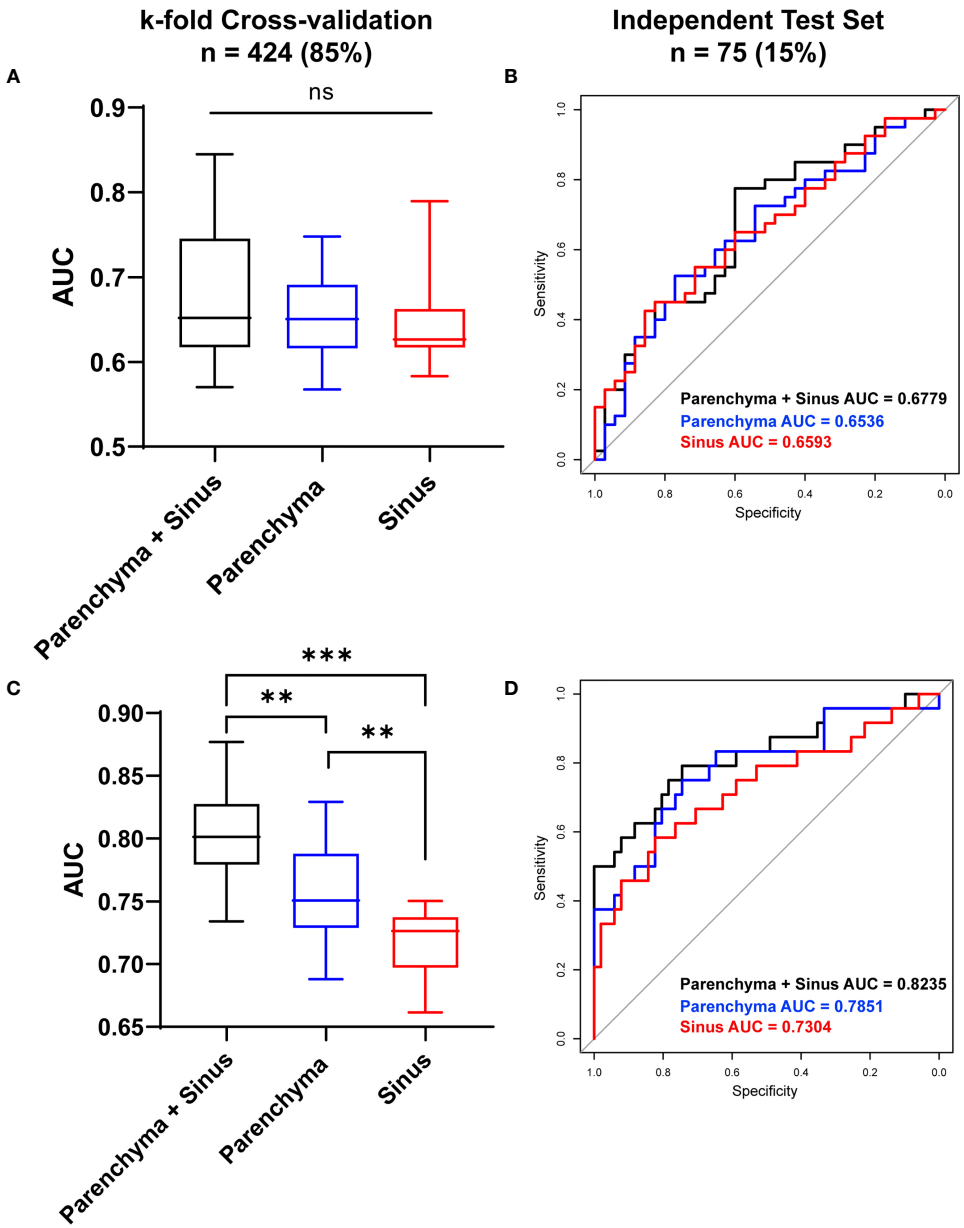


FIGURE 5 | Diagnostic Performance of ultrasound-based radiomics to stratify DKD patients. The diagnostic performance to differentiate DKD and T2DM patient in cross-validation datasets **(A)** and in independent test set **(B)**. The diagnostic performance to differentiate high (≥ stage III) and low (≤ stage II) DKD stages in Cross-validation datasets **(C)** and in independent test set **(D)**. AUC, Area under curve; *P < 0.05, **P < 0.01, ***P < 0.001; ns, no significance.

of radiomics technology. Therefore, in this retrospective multicenter study, we developed a DL-based automatic segmentation, radiomics technology to evaluate the diagnostic performance of radiomics technology for DKD patients and evaluated its potential for clinical application.

Moreover, in underdeveloped regions or primary clinics, it may be costly and challenging to train or recruit experienced doctors to fulfill the large medical demand. One of the solutions is fifth generation communication technology, which can achieve remote medical systems by connecting experienced doctors with patients online. In addition, the DL-based automatic segmentation, radiomics technology can perform the role of an intelligent machine doctor, which is portable if configured using a handheld ultrasound device and can quickly

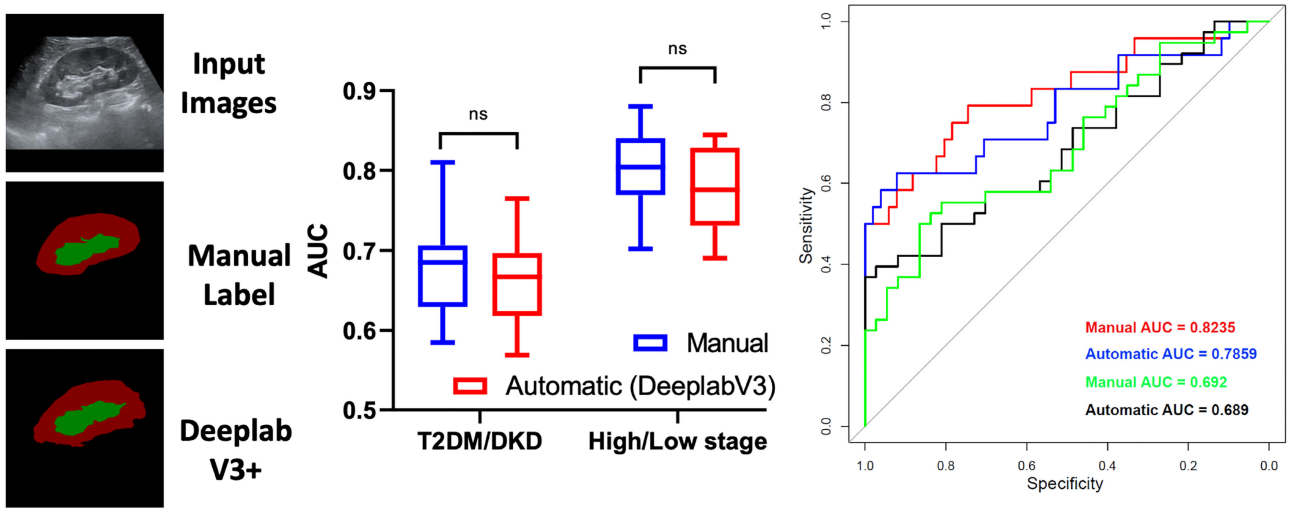


FIGURE 6 | Diagnostic Performance of deep learning-based automatic segmentation, radiomics for diabetic kidney disease. ns, no significance.

and effectively determine the results. This work demonstrates significant potential for achieving automatic labor-free diagnosis and follow-up of DKD, and is our further research focus.

However, this study has certain limitations. First, the retrospective nature of the study may have influenced the accuracy of the diagnostic performance to a certain extent. Second, the heterogeneity among the three medical centers may affect the degree of accuracy. Third, patients with stage I DKD were unavailable in this study. The clinical definition of these patients is primarily defined by pathologic findings that are difficult to access in clinical practice. This resulted in a mixture of T2DM patients without kidney function injury and DKD stage I patients in the dataset of T2DM participants.

In summary, we first verified the robustness of the automatic segmentation method using the DeepLabV3+ network. This result is supported by the research work of Yin et al. (20) the DL-based classification network can achieve good segmentation of the kidney. Moreover, we identified that the renal parenchyma and sinus can provide different information to support the classification model. Accurate anatomical-level segmentation was achieved in this study. The automatic segmentation network achieved superior performance in the segmentation of ultrasonography images with both good and bad ultrasound windows. The Miou and mPA of the automatic segmentation method were high for the independent test set and the datasets from the other independent medical centers. In fact, the DL-based segmentation could reduce the time for hand annotation from 1 h to lower than a few seconds for 100 images, which significantly reduces labor costs. Second, we demonstrated that the ultrasound-based radiomics model achieves a high diagnostic value when differentiating between different DKD stages and has the potential to stratify patients with DKD. The diagnostic performance of artificial intelligence (AI) technology has been supported by the work of Sudharson et al. (21) for certain kidney disorders. Moreover, the work of Chin-Chi Kuo et al. also supported our result (22) in which the authors reported that the prediction of the eGFR and accuracy (85.6%) by an AI-based model was higher than that by experienced nephrologists (60.3-80.1%). The AUC of ROC in our study was 0.803 ± 0.037, which is marginally lower than that reported by Chin-Chi Kuo et al. This may be due to the different clinical definitions of DKD and CKD and the relatively smaller number of patients with severe kidney dysfunctions (eGFR < 30 mL/min/1.73 m^2) in our study (22). Li et al. reported that 3D ultrasound also has potential value in the diagnosis of diabetic nephropathy (DN) and may act as an auxiliary diagnosis for DN (23), which suggests that 3D ultrasound radiomics can be considered in future studies.

CONCLUSION

In this study, we developed a DL-based automatic segmentation, radiomics technology to stratify DKD patients, which could reduce the time for hand annotation from few hours to less than a few seconds for 100 images and could achieve satisfactory diagnostic performance in the diagnosis and follow-up of DKD patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

This study involving human participants was reviewed and approved by the ethics committee of The Second Affiliated Hospital of Zhejiang University School of Medicine. Written informed consent for participation was not required for this study in accordance with the national legislation and institutional requirements.

AUTHOR CONTRIBUTIONS

PH, XJ, and QC designed the study, participated in the supervision and coordination of the study; JC, PJ, YS, and LF conceived and designed the study; JL, HC, LX, FQ, ZC, JS, YaZ, WX, CC, YanZ, JY, and CZ collect the data; JC, PJ, YS, and LF analyzed the data; PH, XJ, QC, JC, PJ, YS, and LF contributed to the writing, review and revision of the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (NO. 82030048, 82102191, 82001818, and 82102052), Key Research and Development Program of Zhejiang Province (NO. 2019C03077), Natural Science Foundation of Zhejiang Province (NO. LQ20H180009, Y16H180019, LQ19H180004 and LQ21H180007).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.876967/full#supplementary-material

REFERENCES

1. Gil CL, Hooker E, Larrivée B. Diabetic Kidney Disease, Endothelial Damage, and Podocyte-Endothelial Crosstalk. *Kidney Med* (2021) 3(1):105–15. doi: 10.1016/j.xkme.2020.10.005
2. Reutens AT. Epidemiology of Diabetic Kidney Disease. *Med Clin North Am* (2013) 97(1):1–18. doi: 10.1016/j.mcna.2012.10.001
3. Zhang L, Long J, Jiang W, Shi Y, He X, Zhou Z, et al. Trends in Chronic Kidney Disease in China. *N Engl J Med* (2016) 375(9):905–6. doi: 10.1056/NEJMc1602469
4. Meola M, Samoni S, Petrucci I. Imaging in Chronic Kidney Disease. *Contrib Nephrol* (2016) 188:69–80. doi: 10.1159/000445469
5. Ilaria P, Anna C, Concetto S, Irene T. Ultrasound and Color Doppler Applications in Chronic Kidney Disease. *J Nephrol* (2018) 31(6):863–79.
6. Bob F, Grosu I, Sporea I, Bota S, Popescu A, Sima A, et al. Ultrasound-Based Shear Wave Elastography in the Assessment of Patients With Diabetic Kidney Disease. *Ultrasound Med Biol* (2017) 43(10):2159–66. doi: 10.1016/j.ultrasmedbio.2017.04.019
7. Wang L, Wu J, Cheng JF, Liu XY, Ma F, Guo LH, et al. Diagnostic Value of Quantitative Contrast-Enhanced Ultrasound (Ceus) for Early Detection of Renal Hyperperfusion in Diabetic Kidney Disease. *J Nephrol* (2015) 28 (6):669–78. doi: 10.1007/s40620-015-0183-3
8. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, et al. Promises and Challenges for the Implementation of Computational Medical Imaging (Radiomics) in Oncology. *Ann Oncol* (2017) 28(6):1191–206. doi: 10.1093/annonc/mdx034
9. Hu HT, Wang Z, Huang XW, Chen SL, Zheng X, Ruan SM, et al. Ultrasound-Based Radiomics Score: A Potential Biomarker for the Prediction of Microvascular Invasion in Hepatocellular Carcinoma. *Eur Radiol* (2019) 29 (6):2890–901. doi: 10.1007/s00330-018-5797-0
10. L-C Chen, Y Zhu, G Papandreou, F Schroff, H Adam eds. *Encoder-Decoder With Atrous Separable Convolution for Semantic Image Segmentation*. Cham: Springer International Publishing (2018) vol 11211. doi: 10.1007/978-3-030-01234-2_49
11. Gosmanov AR, Wall BM, Gosmanova EO. Diagnosis and Treatment of Diabetic Kidney Disease. *Am J Med Sci* (2014) 347(5):406–13. doi: 10.1097/MAJ.0000000000000185
12. Nephrology TCSo. Chinese Guidelines for Diagnosis and Treatment of Diabetic Kidney Disease. *Chin J Nephrol* (2020) 37:3.
13. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* (2017) 77(21):e104–e7. doi: 10.1158/0008-5472.Can-17-0339
14. McEligot AJ, Poynor V, Sharma R, Panangadan A. Logistic Lasso Regression for Dietary Intakes and Breast Cancer. *Nutrients* (2020) 12(9):2652. doi: 10.3390/nu12092652
15. Chen KM, Cofer EM, Zhou J, Troyanskaya OG. Selene: A Pytorch-Based Deep Learning Library for Sequence Data. *Nat Methods* (2019) 16(4):315–8. doi: 10.1038/s41592-019-0360-8
16. Huang P. An Integrated Approach to Ultrasound Imaging in Medicine and Biology. *Bio Integration* (2020) 1(3):105–9. doi: 10.15212/bioi-2020-0036
17. Kelahan LC, Desser TS, Troxell ML, Kamaya A. Ultrasound Assessment of Acute Kidney Injury. *Ultrasound Q* (2019) 35(2):173–80. doi: 10.1097/ruq.0000000000000389
18. Drudi FM, Cantisani V, Granata A, Angelini F, Messineo D, De Felice C, et al. Multiparametric Ultrasound in the Evaluation of Kidney Disease in Elderly. *J Ultrasound* (2020) 23(2):115–26. doi: 10.1007/s40477-019-00390-5
19. Vegar Zubović S, Kristić S, Sefić Pašić I. Relationship Between Ultrasonographically Determined Kidney Volume and Progression of Chronic Kidney Disease. *Med Glas (Zenica)* (2016) 13(2):90–4. doi: 10.17392/852-16
20. Yin S, Peng Q, Li H, Zhang Z, You X, Fischer K, et al. Automatic Kidney Segmentation in Ultrasound Images Using Subsequent Boundary Distance Regression and Pixelwise Classification Networks. *Med Image Anal* (2020) 60:101602. doi: 10.1016/j.media.2019.101602
21. Sudharson S, Kokil P. An Ensemble of Deep Neural Networks for Kidney Ultrasound Image Classification. *Comput Methods Programs BioMed* (2020) 197:105709. doi: 10.1016/j.cmpb.2020.105709
22. Kuo CC, Chang CM, Liu KT, Lin WK, Chiang HY, Chung CW, et al. Automation of the Kidney Function Prediction and Classification Through Ultrasound-Based Kidney Imaging Using Deep Learning. *NPJ Digit Med* (2019) 2:29. doi: 10.1038/s41746-019-0104-2
23. Li N, Wang YR, Tian XQ, Lin L, Liang SY, Li QY, et al. Potential Value of Three-Dimensional Ultrasonography in Diagnosis of Diabetic Nephropathy in Chinese Diabetic Population With Kidney Injury. *BMC Nephrol* (2020) 21 (1):243. doi: 10.1186/s12882-020-01902-w

Conflict of Interest: Author HC was employed by Hangzhou Supor South Ocean Pharmaceutical Co., Ltd.

The remaining 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 06 July 2022
doi: 10.3389/fonc.2022.941349

Machine Learning-Based Models Enhance the Prediction of Prostate Cancer

*Sunmeng Chen[1], Tengteng Jian[1], Changliang Chi[1], Yi Liang[2], Xiao Liang[2], Ying Yu[1], Fengming Jiang[1] and Ji Lu[1]**

[1] Department of Urology, The First Hospital of Jilin University, Changchun, China, [2] School of Business and Management, Jilin University, Changchun, China

Purpose: PSA is currently the most commonly used screening indicator for prostate cancer. However, it has limited specificity for the diagnosis of prostate cancer. We aim to construct machine learning-based models and enhance the prediction of prostate cancer.

Methods: The data of 551 patients who underwent prostate biopsy were retrospectively retrieved and divided into training and test datasets in a 3:1 ratio. We constructed five PCa prediction models with four supervised machine learning algorithms, including tPSA univariate logistic regression (LR), multivariate LR, decision tree (DT), random forest (RF), and support vector machine (SVM). The five prediction models were compared based on model performance metrics, such as the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, calibration curve, and clinical decision curve analysis (DCA).

Results: All five models had good calibration in the training dataset. In the training dataset, the RF, DT, and multivariate LR models showed better discrimination, with AUCs of 1.0, 0.922 and 0.91, respectively, than the tPSA univariate LR and SVM models. In the test dataset, the multivariate LR model exhibited the best discrimination (AUC=0.918). The multivariate LR model and SVM model had better extrapolation and generalizability, with little change in performance between the training and test datasets. Compared with the DCA curves of the tPSA LR model, the other four models exhibited better net clinical benefits.

Conclusion: The results of the current retrospective study suggest that machine learning techniques can predict prostate cancer with significantly better AUC, accuracy, and net clinical benefits.

Keywords: prostate cancer, machine learning, prostate-specific antigen, prostate biopsy, prediction models

OPEN ACCESS

Edited by:
Jian Lu,
Peking University Third Hospital, China

Reviewed by:
Clara Cerrato,
University of California, San Diego, United States
Yishuo Wu,
Fudan University, China

****Correspondence:***
Ji Lu
lu_ji@jlu.edu.cn

Specialty section:
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

Received: *11 May 2022*
Accepted: *13 June 2022*
Published: *06 July 2022*

Citation:
Chen S, Jian T, Chi C, Liang Y, Liang X, Yu Y, Jiang F and Lu J (2022) Machine Learning-Based Models Enhance the Prediction of Prostate Cancer. Front. Oncol. 12:941349. doi: 10.3389/fonc.2022.941349

INTRODUCTION

Prostate cancer (PCa) is the second leading malignancy in men and the fifth leading cause of cancer mortality in men worldwide (1). Although PSA is still the most commonly used screening tool for prostate cancer, it has been controversial in recent decades (2). It is suggested that PSA screening improves the detection rate of localized and less aggressive prostate cancers but also reduces the

proportions of advanced PCa and PCa-specific mortality (3–5). However, due to the obvious overlap of PSA levels in various conditions, such as benign prostatic hyperplasia, prostatitis, and prostate cancer, the specificity of PSA screening is low, which leads to a plethora of unrelated diseases for prostate biopsy (2). These unnecessary prostate biopsies result in not only a significant waste of medical resources but also an increased incidence of sepsis, which can be life-threatening to patients (6). Therefore, there is a need for a new convenient method to improve the diagnostic ability of PCa.

Machine learning is a branch of artificial intelligence (AI) in which machines are programmed to learn patterns from data, and the learning itself is based on a set of mathematical rules and statistical assumptions. It is widely used in biology because of its enormous advantages in dealing with large datasets (7, 8). It has also been rapidly developed and applied in the medical field, especially in the construction of predictive models (9). Therefore, using machine learning to construct PCa prediction models would be a feasible and promising approach.

In this study, we constructed generalizable machine learning predictive models to improve the accuracy of PCa risk assessment by using objective parameters present in electronic medical records and then evaluated their performance.

METHODS

Data Sources

A total of 789 male patients in the First Hospital of Jilin University who underwent transrectal ultrasound-guided prostate biopsy from January 2013 to January 2021 were included. Indications for prostate biopsy included serum tPSA >4 ng/ml, abnormal digital rectal exam (DRE), or imaging findings suggestive of suspected prostate cancer. All patients underwent systematic biopsy with 10-12 cores. Patients with one of the following criteria were excluded from the study: taking medications that could affect serum PSA levels, unclear results of the prostate biopsy, and significant abnormal values or missing data. A final total of 551 patients were included in the study. All patient data were collected through electronic medical records, including age, BMI, hypertension, diabetes, total PSA (tPSA), free PSA (fPSA), the ratio of serum fPSA to tPSA (f/tPSA), prostate volume (PV), PSA density (PSAD), neutrophil-to-lymphocyte ratio (NLR), and pathology reports of prostate biopsy. All examinations were completed within one week before prostate biopsy. PSAD is the ratio of tPSA to PV. The calculation of PV was calculated by the following formula: maximal transverse diameter × maximal anterior-posterior diameter × maximal superior-inferior diameter × 0.52.

Model Development

R software (version 4.1.4, https://www.rproject.org/) was used to develop machine learning models. A total of five prediction models were constructed by dividing the data into a training dataset and a test dataset at a ratio of 3:1. The pathologic type of PCa or other benign disease was used as the dichotomous variable, and other variables were all used as continuous variables. For logistic regression (LR) model construction, the best predictive variables were first screened in the training dataset using stepwise regression, association plots between the variables were made to understand the magnitude of the association between the variables, and the presence or absence of collinearity between variables was judged according to the variance inflation factor (VIF). Then, LR models were constructed using the "lrm" function in the "rms" package. For the decision tree (DT) model, we used the "rpart" package for training, using two hyperparameters, the complexity parameters cp and spilt. The initial cp value was set to 0.001, and then the best cp value was found and pruned based on the best cp value. The input variables were obtained through the selection of important features, and the best DT model was then output. The random forest (RF) model screened the optimal input variables by significant feature selection. The RF model was trained using the "randomForest" package in R software, using two hyperparameters, ntree and mtry, which were set at 500 and 6, respectively. The support vector machine (SVM) model was filtered by the "caret" package for important features. Training was performed using the "e1071" package, using a Gaussian kernel function and setting the two hyperparameters, cost and gamma, to 1 and 0.1, respectively.

Model Performance Evaluation

The performance of the developed models was validated using a test dataset in a process that was completely independent of the algorithm training. The performance of the five models was then evaluated by comparing four metrics: the receiver operating characteristic (ROC) curve and its corresponding area under the curve (AUC) and the accuracy, sensitivity, and specificity. The calibration curve was used to evaluate the calibration of the model, the Brier score was used to assess the calibration, the Hosmer–Lemeshow goodness-of-fit test was used to judge whether there was a significant difference between the observed and predicted values, and clinical decision curve analysis (DCA) was used to assess the net benefit of the model.

Data Analysis

For comparative analysis between two samples, Student's t test was used for normally distributed continuous variables, and the Mann–Whitney U test was used for categorical variables with nonnormal continuous variables. Continuous variables in the data were expressed as medians and IQRs or means and SDs, categorical variables were expressed as frequencies and percentages, and the bilateral significance level for the left-right test was set at 5% ($p<0.05$).

RESULTS

Baseline Patient Characteristics

Table 1 shows the baseline characteristics of the patients. A total of 302 (54.8%) of the 551 patients were diagnosed with PCa. The PCa detection rate in patients ≥65 years old was higher than patients <65. The mean levels of tPSA, fPSA, and PSAD were significantly higher in the PCa group than in the non-PCa group. When tPSA>4 ng/ml, the PCa detection rate increased with increasing tPSA. In the subgroups of 4<tPSA<10, 10≤tPSA<20, 20≤tPSA<100, and tPSA≥100, the detection rates were 18.6%,

TABLE 1 | Characteristics of Patients, stratified by biopsy outcomes.

	Total (n=551)	Non-PCa (n=249)	PCa (n=302)	PCa detection (%)	p value
Age,year(n,%)					< 0.001
<65	157 (28.5)	96 (38.6)	61 (20.2)	38.9	
≥65	394 (71.5)	153 (61.4)	241 (79.8)	61.2	
BMI,kg/m2(SD)	23.6 ± 3.2	24.0 ± 3.0	23.3 ± 3.4		0.013
Hypertension,(n,%)					0.059
No	416 (75.5)	178 (71.5)	238 (78.8)	57.2	
Yes	135 (24.5)	71 (28.5)	64 (21.2)	47.4	
Diabetes,(n,%)					0.603
No	511 (92.7)	233 (93.6)	278 (92.1)	54.4	
Yes	40 (7.3)	16 (6.4)	24 (7.9)	60.0	
Neutrophil count, 10^9/L(IQR)	3.7 (2.9-4.8)	4.0 (3.1-5.0)	3.7 (2.8-4.6)		0.024
Lymphocyte count, 10^9/L(SD)	1.8 ± 0.7	1.8 ± 0.6	1.8 ± 0.7		0.798
NLR,(IQR)	2.1 (1.5-3.0)	2.1 (1.5-3.3)	2.0 (1.5-2.9)		0.150
Median tPSA, ng/ml(IQR)	32.0 (14.5-100.0)	16.2 (10.6-26.1)	88.8 (34.0-123.2)		< 0.001
tPSA, ng/ml,(n,%)					< 0.001
≤4	13 (2.4)	10 (4)	3 (1)	23.1	
4<tPSA<10	59 (10.7)	48 (19.3)	11 (3.6)	18.6	
10≤tPSA<20	122 (22.1)	90 (36.1)	32 (10.6)	26.2	
20≤tPSA<100	212 (38.5)	97 (39)	115 (38.1)	54.2	
≥100	145 (26.3)	4 (1.6)	141 (46.7)	97.2	
fPSA,ng/ml(SD)	13.6 ± 31.3	3.3 ± 5.8	22.2 ± 39.9		< 0.001
f/tPSA,(n,%)					0.280
<0.16	390 (70.8)	170 (68.3)	220 (72.8)	56.4	
≥0.16	161 (29.2)	79 (31.7)	82 (27.2)	50.9	
PV,cm^3(IQR)	55.3 (37.2-79.0)	65.2 (44.7-91.5)	48.9 (34.3-70.3)		< 0.001
PSAD,ng/ml/cm^3(IQR)	0.6 (0.3-1.6)	0.3 (0.2-0.4)	1.5 (0.7-2.7)		< 0.001

PV, prostate volume; PSAD, prostate-specific antigen density; BMI, body mass index; NLR, neutrophil-to-lymphocyte ratio; tPSA, total prostate specific antigen; fPSA, free prostate specific antigen; PCa, prostate cancer; Non-PCa, non-prostate cancer.

26.2%, 54.2%, and 97.2%, respectively. The mean neutrophil count, PV, and BMI were lower in the PCa group than in the non-PCa group. No significant differences in other variables were found between two groups.

LR Algorithm-Based PCa Prediction Model

First, the tPSA univariate LR model was constructed in the training dataset by including only one single factor, tPSA. This model showed that tPSA was positively correlated with the diagnosis of PCa (coefficient=0.034). Then, a multivariate LR model was constructed by including all variables through a stepwise regression method. When the VIC value reached the minimum value, a total of seven best predictive variables were selected, including age, tPSA, fPSA, PV, NLR, peripheral blood neutrophil count and lymphocyte count. The two LR models are shown in **Supplementary Table 1**. In the multivariate LR model, age, tPSA, and fPSA were positively correlated with PCa, while PV and neutrophil count were negatively correlated with PCa (**Supplementary Figure 1**). In addition, there was a significant association between peripheral blood neutrophil count and NLR, tPSA and fPSA, suggesting their respective possible collinearity. The variance inflation factor (VIF) was subsequently calculated for verification, and all VIF values were less than 5, indicating no collinearity between any of the variables.

DT Algorithm-Based PCa Prediction Model

The optimal cp value for the DT model was 0.008. Based on the corresponding ranking of important features, the final input variables for the DT model were age, tPSA, fPSA, PV, PSAD, NLR, f/tPSA, and biopsy results (**Supplementary Figure 2A**). The process and results of model classification are shown in **Supplementary Figure 2B**. This model correctly classified 87.7% (363/414) of the cases in the training dataset.

RF Algorithm-Based PCa Prediction Model

After ranking the important features of the RF model, seven features with the highest predictive accuracy were selected as the input features, including age, tPSA, fPSA, PV, PSAD, NLR, and peripheral blood neutrophil count (**Supplementary Figure 3A**). The error of the model gradually decreased as the number of decision trees increased, and the minimum error value of the RF model was reached when the number of decision trees was 313 (**Supplementary Figure 3B**).

SVM Algorithm-Based PCa Prediction Model

The best SVM model was screened by the "rfe" function in the caret package using a 10-fold cross-validation method. The number of variables was screened one by one from 1 to 12, and the best model was obtained when the number of variables was 5. The input variables at this time were age, PSAD, tPSA, fPSA, and PV.

Performance of the Developed Models

In the training dataset, the RF and DT models performed particularly well in differentiation with the AUCs of 1.0 and 0.922, respectively (**Table 2**). The model with the lowest AUC was the tPSA univariate LR model (0.842). The calibration in all five models was very good, which suggested that the predicted values of the models were in high agreement with the actual

TABLE 2 | Diagnostic performance of different machine learning models.

Outcome	Dataset	tPSAlogisticregression	Multivariate logisticregression	Decision Tree	Random Forest	SupportVectorMachine
AUC	Training	0.842	0.910	0.922	1.00	0.884
	Test	0.846	0.918	0.886	0.898	0.895
Sensitivity(%)	Training	69.2	70.5	91.2	100	86.8
	Test	63.9	88.0	86.7	84.0	86.7
Specificity(%)	Training	88.8	95.2	83.4	100	70.6
	Test	93.2	87.1	69.4	79.0	85.5
Accuracy(%)	Training	78.0	81.6	87.7	100	79.5
	Test	77.1	87.6	78.8	81.8	86.1

values (**Figure 1A**). The Brier scores of the multivariate LR, DT, RF, SVM and tPSA LR models were 0.119, 0.122, 0.121, 0.118 and 0.154, respectively. The p values for the Hosmer–Lemeshow test were all greater than 0.05, indicating that there was no statistical bias in the near-perfect fit between the predicted and actual values. Therefore, the models we constructed were valid and reliable.

All characteristics of samples were comparable in the training and test datasets (**Supplementary Table 2**). In the test dataset, the specificity of the tPSA LR model was highest, reaching 93.2%; however, the sensitivity and accuracy were relatively low, at 63.9% and 77.1%, respectively (**Table 2**). The sensitivity and accuracy of the multivariate LR model were improved significantly in the test dataset compared with the training dataset, although there was a slight decrease in specificity. The diagnostic performance of the SVM model was not outstanding in either the training or the test dataset, but most outcomes were improved in the test dataset. Thus, the multivariate LR model exhibited the best discrimination, and the extrapolation and generalization abilities of the multivariate LR and SVM models were relatively strong. In contrast, although the DT and RF models performed well in the training dataset, their performances in the test dataset decreased significantly. The corresponding ROC curves for the five models are shown in **Figure 1B**. To further evaluate the potential clinical benefits of these prediction models, we performed DCA curves using the test dataset (**Figure 1C**). All models demonstrated significant net benefits. Compared with the tPSA LR model, the other four models showed significantly higher net clinical benefits.

To evaluate the performance of constructed models in the subgroups of PSA of 4-10 ng/ml and 4-20 ng/ml, we used these two subgroups as the test datasets, and evaluated their AUCs, sensitivities, specificities, and accuracies (**Supplementary Tables 3, 4**). The diagnostic performances of the RF algorithm-based machine learning prediction model showed best in both subgroups. The AUCs were 0.856 and 0.94 respectively. Although the sensitivity decreased compared with it in the training dataset, the specificity and accuracy were still relatively high. The performance of other algorithms-based developed models were not outstanding.

DISCUSSION

In the past decade, PSA has been widely used as the most valuable diagnostic and prognostic marker for PCa (10). However, some studies have shown that less than 30% of men with PSA levels in the gray zone (4-10 ng/ml) have pathologically confirmed PCa, indicating that a large proportion of patients have undergone unnecessary biopsies and been overdiagnosed (11). In our study, the detection rate of PCa with tPSA in the gray zone was only 18.6%. Similar detection rates were reported in other studies (12–15). Even if PSA is not in the gray zone, for example, between 10 and 20 ng/ml, the detection rate in our study was only 26.2%. In recent years, DRE, PSAD, PSAV, 4Kscore, f/tPSA ratio, prostate health index (PHI) and age-specific PSA have been proposed

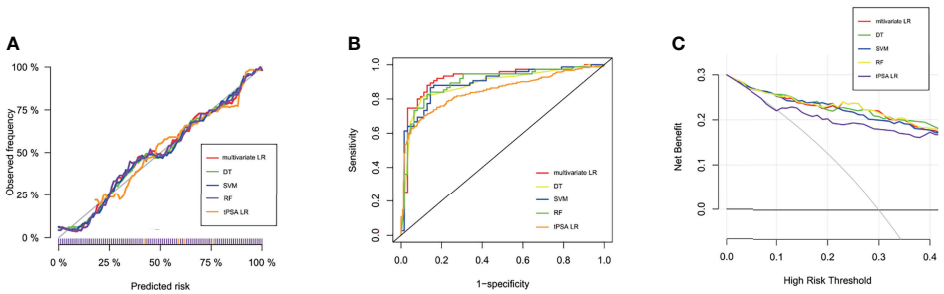


FIGURE 1 | Performance of machine learning models. **(A)** Calibration curves of five prediction models in the training dataset. The predicted probabilities are plotted on the X-axis, and the actual probabilities are plotted on the Y-axis. **(B)** The ROC curves of the tPSA LR, multivariate LR, DT, RF, and SVM models in the test dataset. **(C)** Clinical decision curve analysis (DCA) of the five models in the test dataset.

as predictors for PCa (3, 16–18). However, it may be difficult to achieve good predictive results with any single factor.

In this study, we constructed prediction models of PCa based on machine learning algorithms. Four algorithms were used, and a total of five models were constructed. First, we constructed a univariate logistic regression model using tPSA. As shown in **Table 2**, in the test dataset, although the specificity of the tPSA LR model reached 0.932, its sensitivity was decreased to only 0.639. The AUC of this model was only 0.846, which was significantly lower than that of the other models, showing the limitations of using tPSA alone as a predictor of PCa. Among them, the multivariate LR model had higher specificity, sensitivity, accuracy, and AUC in the test dataset, showing good predictive ability. Therefore, the shortcomings of low sensitivity and accuracy of the tPSA LR model were complemented very well by the inclusion of more variables. In our study, the multivariate LR model also had outstanding extrapolation and generalization ability due to the small number of changes between the training and test datasets. As a model similar to traditional statistical analysis methods, the results of the LR model had strong interpretability, which could help clinicians predict PCa based on relevant factors. The output of the DT model was similar to the clinical pathway. It is a clinician- and patient-friendly model and has strong clinical operability. Anyone can follow the predicted model from the root node to the leaf node to make decisions. However, in our study, the performance of the constructed DT model decreased significantly when it was validated in the test dataset and had very low specificity. The RF and SVM models had average diagnostic performance in our study, and their "black box" style reduced the clinical interpretability slightly.

The PCa detection rate in patients with PSA<20 ng/ml was relatively low in our study. Thus it is important to predict PCa in this population. The RF model performed best in both subgroups PSA of 4-10 ng/ml and 4-20 ng/ml. This suggested that the machine learning models we constructed based on overall population might also be applicable in patients with PSA ranged 4-10 or 4-20 ng/ml. In the training dataset, the RF model outperformed other models with the AUC of 1.0, while its performance decreased significantly in the test dataset. But in these two PSA subgroups, about 72% of samples overlapped with the samples in the training dataset, and were involved in the construction of models. That may lead to the outperformance of RF model rather than other models in PSA subgroups. Future study should focus on this population and develop more accurate machine learning models.

In recent years, some studies on the prediction of PCa by machine learning models have been published. In the study of Peter Ka-Fung Chiu et al., four variables, PSA, DRE, PV and transrectal ultrasound findings, were included, and SVM, LR, and RF models were constructed. All models were shown to have better prediction for PCa and clinically significant PCa than PSA and PSAD alone (19). Similarly, Nitta et al. suggested that compared to the AUCs of the PSA level, PSAD, and PSAV alone, the AUCs of artificial neural network (ANN), RF, SVM machine learning models were all improved when age, PSA level, PV, and white blood cell count in urinalysis were incorporated (20).In a study including patients with tPSA<10 ng/ml, a PSA-based machine learning model was constructed based on dense neural network with an AUC of 0.72, which was improved compared to PSA alone, age, fPSA and f/tPSA alone (21). In another study, multiparametric MRI (mpMRI) combined with other characteristics of patients was included to construct a machine learning model. The SVM and RF yielded similar diagnostic accuracy and net benefit and spared more biopsies at 95% sensitivity for the detection of clinically significant PCa compared with logistic regression (15).

Recently, a meta-analysis showed that the performance of selectMDx test in urine was comparable to that of mpMRI with regards to PCa detection. The AUC of selectMDx only was 0.854, and the AUC of one or both positive finding with selectMDx and/or mpMRI could reach 0.909. However, the multivariate LR model in our study still shows strength with AUC of 0.918 in test dataset (22). The integrative machine learning model was constructed to predict negative prostate biopsy utilizing both radiomics and clinical features. Although that model got high performance with negative predictive value of 98.3%, the AUC, sensitivity, and specificity were 0.798, 83.3%, and 75.2%, respectively, which were relatively lower compared with those in our models (23). Considering that all variables in our models are objective indicators, reducing possible errors of manual evaluation, we believe that the advantages of our models are more obvious.

Although machine learning-based models for PCa prediction have been constructed and validated, there are several limitations in our study. First, our study is retrospective in nature and may be potentially biased, and the sample size may not be adequate for some machine learning algorithms. Second, both the training and test datasets were from the same hospital, so further external validation at other centers is needed to confirm the findings. Third, some important factors or variables were not included in our models. For example, it has been shown that mpMRI provides more imaging information than conventional ultrasound, not only improving the detection of PCa but also helping to distinguish clinically significant PCa (24). Limited by insufficient data, mpMRI and its PI-RADS data were not included in our study. It is hoped that incorporating mpMRI into machine learning models may help to further improve the diagnostic performance of models in the future.

CONCLUSION

In conclusion, by retrieving electronic medical records, we developed, validated, and compared machine learning models to predict PCa in the biopsy population. All models showed clinical benefits based on DCA. Multivariate LR, DT, RF, and SVM models were better than tPSA univariate LR. Among these models, multivariate LR performed best, with an AUC of 0.918 in the test dataset. Constructing machine learning-based models and predicting PCa is feasible. This could enhance the detection of PCa and help to avoid unnecessary prostate biopsy.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SC: Data analysis, Manuscript writing/editing. TJ: Data collection. CC: Data collection. YL: Data analysis. XL: Data analysis. YY: Data collection. FJ: Project development, supervision. JL: Project development, supervision, manuscript writing/editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Science and Technology Development Project of Jilin Province, China (20200201315JC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.941349/full#supplementary-material

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: Cancer J Clin* (2021) 71 (3):209–49. doi: 10.3322/caac.21660
2. Eldred-Evans D, Tam H, Sokhi H, Padhani AR, Winkler M, Ahmed HU. Rethinking Prostate Cancer Screening: Could Mri Be an Alternative Screening Test? *Nat Rev Urol* (2020) 17(9):526–39. doi: 10.1038/s41585-020-0356-2
3. Catalona WJ. Prostate Cancer Screening. *Med Clinics North America* (2018) 102(2):199–214. doi: 10.1016/j.mcna.2017.11.001
4. Tsodikov A, Gulati R, Heijnsdijk EAM, Pinsky PF, Moss SM, Qiu S, et al. Reconciling the Effects of Screening on Prostate Cancer Mortality in the Erspc and Plco Trials. *Ann Internal Med* (2017) 167(7):449–55. doi: 10.7326/m16-2586
5. Fenton JJ, Weyrich MS, Durbin S, Liu Y, Bang H, Melnikow J. Prostate-Specific Antigen-Based Screening for Prostate Cancer: Evidence Report and Systematic Review for the Us Preventive Services Task Force. *Jama* (2018) 319 (18):1914–31. doi: 10.1001/jama.2018.3712
6. Tan GH, Nason G, Ajib K, Woon DTS, Herrera-Caceres J, Alhunaidi O, et al. Smarter Screening for Prostate Cancer. *World J Urol* (2019) 37(6):991–9. doi: 10.1007/s00345-019-02719-5
7. Greener JG, Kandathil SM, Moffat L, Jones DT. A Guide to Machine Learning for Biologists. *Nat Rev Mol Cell Biol* (2022) 23(1):40–55. doi: 10.1038/s41580-021-00407-0
8. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. *Cell* (2018) 173 (7):1581–92. doi: 10.1016/j.cell.2018.05.015
9. Schwalbe N, Wahl B. Artificial Intelligence and the Future of Global Health. *Lancet* (2020) 395(10236):1579–86. doi: 10.1016/s0140-6736(20)30226-9
10. Matlaga BR, Eskew LA, McCullough DL. Prostate Biopsy: Indications and Technique. *J Urol* (2003) 169(1):12–9. doi: 10.1097/01.ju.0000041828.84343.53
11. Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, et al. Measurement of Prostate-Specific Antigen in Serum as a Screening Test for Prostate Cancer. *New Engl J Med* (1991) 324(17):1156–61. doi: 10.1056/nejm199104253241702
12. Lee A, Chia SJ. Contemporary Outcomes in the Detection Of Prostate Cancer Using Transrectal Ultrasound-Guided 12-Core Biopsy in Singaporean Men With Elevated Prostate Specific Antigen and/or Abnormal Digital Rectal Examination. *Asian J Urol* (2015) 2(4):187–93. doi: 10.1016/j.ajur.2015.08.003
13. Seo HK, Chung MK, Ryu SB, Lee KH. Detection Rate of Prostate Cancer According to Prostate-Specific Antigen and Digital Rectal Examination in Korean Men: A Nationwide Multicenter Study. *Urology* (2007) 70(6):1109–12. doi: 10.1016/j.urology.2007.07.052
14. Matsumoto K, Satoh T, Egawa S, Shimura S, Kuwao S, Baba S. Efficacy and Morbidity of Transrectal Ultrasound-Guided 12-Core Biopsy for Detection of Prostate Cancer in Japanese Men. *Int J Urol* (2005) 12(4):353–60. doi: 10.1111/j.1442-2042.2005.01058.x
15. Yu S, Tao J, Dong B, Fan Y, Du H, Deng H, et al. Development and Head-To-Head Comparison of Machine-Learning Models to Identify Patients Requiring Prostate Biopsy. *BMC Urol* (2021) 21(1):80. doi: 10.1186/s12894-021-00849-w
16. Moradi A, Srinivasan S, Clements J, Batra J. Beyond the Biomarker Role: Prostate-Specific Antigen (Psa) in the Prostate Cancer Microenvironment. *Cancer Metastasis Rev* (2019) 38(3):333–46. doi: 10.1007/s10555-019-09815-3
17. Catalona WJ, Partin AW, Slawin KM, Brawer MK, Flanigan RC, Patel A, et al. Use of the Percentage of Free Prostate-Specific Antigen to Enhance Differentiation of Prostate Cancer From Benign Prostatic Disease: A Prospective Multicenter Clinical Trial. *Jama* (1998) 279(19):1542–7. doi: 10.1001/jama.279.19.1542
18. Carter HB, Ferrucci L, Kettermann A, Landis P, Wright EJ, Epstein JI, et al. Detection of Life-Threatening Prostate Cancer With Prostate-Specific Antigen Velocity During a Window of Curability. *J Natl Cancer Inst* (2006) 98 (21):1521–7. doi: 10.1093/jnci/djj410
19. Chiu PK, Shen X, Wang G, Ho CL, Leung CH, Ng CF, et al. Enhancement of Prostate Cancer Diagnosis by Machine Learning Techniques: An Algorithm Development and Validation Study. *Prostate Cancer Prostatic Dis* (2021). doi: 10.1038/s41391-021-00429-x
20. Nitta S, Tsutsumi M, Sakka S, Endo T, Hashimoto K, Hasegawa M, et al. Machine Learning Methods Can More Efficiently Predict Prostate Cancer Compared With Prostate-Specific Antigen Density and Prostate-Specific Antigen Velocity. *Prostate Int* (2019) 7(3):114–8. doi: 10.1016/j.prnil.2019.01.001
21. Perera M, Mirchandani R, Papa N, Breemer G, Effeindzourou A, Smith L, et al. Psa-Based Machine Learning Model Improves Prostate Cancer Risk Stratification in a Screening Population. *World J Urol* (2021) 39(6):1897–902. doi: 10.1007/s00345-020-03392-9
22. Sari Motlagh R, Yanagisawa T, Kawada T, Laukhtina E, Rajwa P, Aydh A, et al. Accuracy of Selectmdx Compared to Mpmri in the Diagnosis of Prostate Cancer: A Systematic Review and Diagnostic Meta-Analysis. *Prostate Cancer Prostatic Dis* (2022) 25(2):187–98. doi: 10.1038/s41391-022-00538-1
23. Zheng H, Miao Q, Liu Y, Raman SS, Scalzo F, Sung K. Integrative Machine Learning Prediction of Prostate Biopsy Results From Negative Multiparametric Mri. *J Magn Reson Imaging* (2022) 55(1):100–10. doi: 10.1002/jmri.27793
24. Stabile A, Giganti F, Rosenkrantz AB, Taneja SS, Villeirs G, Gill IS, et al. Multiparametric Mri for Prostate Cancer Diagnosis: Current Status and Future Directions. *Nat Rev Urol* (2020) 17(1):41–61. doi: 10.1038/s41585-019-0212-4

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

TYPE Mini Review
PUBLISHED 17 August 2022
DOI 10.3389/fonc.2022.990176

OPEN ACCESS

EDITED BY
Xiongbing Zu,
Xiangya Hospital, Central South University, China

REVIEWED BY
Octavian Sabin Tataru,
Sciences and Technology of Târgu Mureç, Romania

*CORRESPONDENCE
Fan Lin
foxetfoxet@gmail.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 09 July 2022
ACCEPTED 27 July 2022
PUBLISHED 17 August 2022

CITATION
Huang X, Wang X, Lan X, Deng J, Lei Y and Lin F (2022) The role of radiomics with machine learning in the prediction of muscle-invasive bladder cancer: A mini review.
Front. Oncol. 12:990176.
doi: 10.3389/fonc.2022.990176

The role of radiomics with machine learning in the prediction of muscle-invasive bladder cancer: A mini review

Xiaodan Huang†, Xiangyu Wang†, Xinxin Lan, Jinhuan Deng, Yi Lei and Fan Lin*

Department of Radiology, The First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, China

Bladder cancer is a common malignant tumor in the urinary system. Depending on whether bladder cancer invades muscle tissue, it is classified into non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). It is crucial to accurately diagnose the muscle invasion of bladder cancer for its clinical management. Although imaging modalities such as CT and multiparametric MRI play an important role in this regard, radiomics has shown great potential with the development and innovation of precision medicine. It features outstanding advantages such as non-invasive and high efficiency, and takes on important significance in tumor assessment and laor liberation. In this article, we provide an overview of radiomics in the prediction of muscle-invasive bladder cancer and reflect on its future trends and challenges.

KEYWORDS

bladder cancer, radiomics, machine learning, muscle-invasive, CT, MRI

1 Introduction

Bladder cancer (BC) is the second most common cancer among urological malignancies, with an estimated 573,200 people diagnosed with BC worldwide in 2020 (1). The rates of bladder cancer increase with age. The risk of BC is multifactorial, with smoking (2) being the most important risk factor. Uroepithelial carcinoma accounts for approximately 90% of bladder cancer cases and typically presents as multifocal and recurrent; other subtypes are squamous cell carcinoma (6-8%) and adenocarcinoma (3).

Determining the invasion of the tumor into the muscle layer of the bladder wall is probably the most critical step in clinical management, as it directly affects the patient's treatment strategy. Bladder cancers are classified into non-muscle-invasive bladder cancer (NMIBC) (≤ T1 stage) and muscle-invasive bladder cancer (MIBC) (≥

T2 stage) according to whether they invade muscle tissue or not. NMIBC is mostly in the early stages of the disease, with a 5-year probability of recurrence and progression of 78% and 45%, respectively (4), while MIBC has a poor prognosis, with approximately 50% (5) of patients developing metastases within 2 years after radical cystectomy(RC). NMIBC is usually treated by transurethral resection of bladder tumors (TURBT) with or without intravesical chemotherapy (6). Whereas MIBC is usually treated with radical cystectomy (RC), radiotherapy, chemotherapy, or combination therapy (5). Currently, pathological examination of TURBT specimens is the gold standard for identification of MIBC. However, according to previous studies, the error rate is about 20-80% due to problems such as differences in resection (7). Even though the error rate can be reduced by repeating TURBT, underestimation of staging and delayed treatment of the condition may lead to disease progression and worse prognosis, and this invasive operation also carries some safetyoperational risks. Faced with the above problems, scholars have searched for an alternative, non-invasive and efficient diagnostic tool to accurately predict muscle-invasive bladder cancer, so they have turned their attention to "radiomics" - a hot and promising diagnostic technology. Radiomics is the extraction and analysis of quantitative imaging features from imaging tools (CT, MRI, PET-CT, etc.) for the development of descriptive and predictive models (8). Machine learning (ML), a branch of artificial intelligence, is a typical approach used in radiomics model generation (9). Through the inferential training of datasets, ML aids in the development of highly accurate and effective predictive models based on radiomics analysis (10). In this paper, we review the current existing research related to our topic, summarize the results of using machine learning to accurately predict muscle-invasive bladder cancer, and reflect on the future directions and challenges of the topic.

2 Search criteria

A comprehensive review of current literature was performed using the PubMed-Medline and Web of Science database up to April 5, 2022 using "bladder cancer", combined with one of the following terms: "radiomics", "machine learning", and "artificial intelligence" in combination with "muscle invasive".

The exclusion criteria for the articles were as follows:

(1) Published in a language other than English.

(2) The purpose of the article study was not to predict muscle invasion of bladder cancer.

(3) The article was not studied with imaging tools.

(4) Reviews, conference abstracts, and editorials were excluded.

The inclusion criteria for the article were as follows:

(1) Background introduction of radiomics, machine learning, deep learning or artificial intelligence and bladder cancer.

(2) The purpose of the article study was to predict muscle invasion of bladder cancer.

(3) The article was studied with imaging tools(CT, MRI, PET-CT, SPECT e.g.).

In accordance with the PRISMA criteria, Figure 1 was included to delineate our article

selection process.

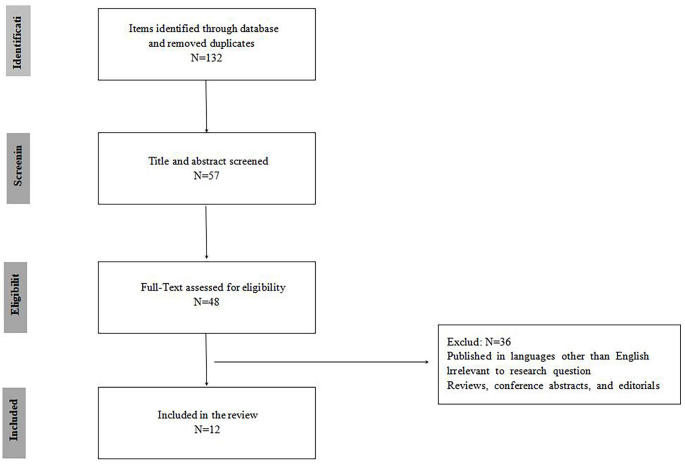


FIGURE 1
PRISMA flowchart of included studies.

3 Results

The final collection of 12 relevant publications found that the first study started in 2017, reflecting the fact that radiomics is a relatively new concept in the field of BC. The literature related to machine learning for predicting muscle-invasive bladder cancer is summarized in Table 1 (11–22). For studies in this field, four were based on enhanced CT and the remaining eight were related to MRI. Only 16.7% (2/12) of the studies were multi-center studies.

4 Discussion

4.1 Traditional diagnostic imaging

In current clinical practice, medical imaging techniques including CT, MRI and other non-invasive and safe diagnostic modalities are increasingly recognized for their performance in predicting muscle invasion and staging of bladder cancer. MRI has mainly been found to play a crucial role in the early localization and

TABLE 1 Studies included in the systematic review.

	Study characteristics			Patient characteristics				Imaging characteristics	
	Author	Year	Study design	Number of cases	Number of selected lesions	Surgical technique	Pathological stage NMIBC: MIBC	Imaging modality	Scanner
1	Xu	2017	Single-center retrospective	78	118	NA	34:84	T2WI	3.0T GE
2	Garapati	2017	Single-center retrospective	76	84	Cystectomy	43:41	CTU	NA
3	Tong	2018	Single-center retrospective	65	65	Cystectomy	31:34	T2WI	1.5-3.0T
4	Xu	2019	Single-center retrospective	54	54	NA	24:30	T2WI、DWI、ADC	3.0T GE
5	Zheng	2019	Single-center retrospective	199	199	RC or TURBT	130:69	T2WI	3.0T MR scanner (Intera Achieva, Philips Medical Systems)
6	Xu	2020	Single-center retrospective	218	218	Both TURBT and RC	131:87	DWI	3.0T MR scanner (Ingenia;Philips Healthcare)
7	Wang	2020	Mult-center retrospective	106	106	RC or partial cystectomy or TURBT	64:42	T2WI、DWI、ADC	3.0T MR system (MAGNETOM Trio, Siemens Healthineers)
8	Hammouda	2021	Single-center retrospective	42		NA		T2WI、DWI、ADC	3.0T Ingenia Philips MRI scanners
9	Zhang	2021	Mult-center retrospective	441	441	RC or TURBT	183 (development):110 (tuning) :73(internal validation) :75(external validation)	Enhanced CT	NA
10	Zheng	2021	Single-center retrospective	185	185	NA	129:56	T2WI、DCE	3.0T MRI scanner(Magnetom Verio: Siemens, Erlangen, Germany)
11	Zhou	2021	Single-center retrospective	100	100	NA	70:30	Enhanced CT	Siemens 64-row spiral CT
12	Cui	2022	Single-center retrospective	327	188	RC or partial cystectomy or TURBT	120:68	CECT	GE Dis covery CT750HD, GE LightSpeed VCT, Philips ICT 256, and Siemens Somatom Definition Flash.

ADC, apparent diffusion coeffificient; CECT, contrast-enhanced computed tomography; CT, computed tomography; CTU, CT Urography; DCE, dynamic contrast enhanced; DWI, diffusion-weighted imaging; MIBC, muscle-invasive bladder cancer; MR, magnetic resonance; MRI, magnetic resonance imaging; NA, not available; NMIBC, non–muscle-invasive bladder cancer; RC, radical cystectomy; TURBT, transurethral resection of bladder tumor; T2WI, T2-weighted imaging.

invasive diagnosis of BC. T2-weighted imaging(T2WI) is able to illustrate detailed structural information of the lesion and bladder wall, thus potentially reflecting the depth of invasion of the bladder wall of BC. The low signal line of the detrusor muscle is interrupted by MIBC, whereas the detrusor muscle is complete in NMIBC. Diffusion Weighted Imaging(DWI) and Apparent Diffusion Coefficient(ADC) have a good ability to reflect signal intensity differences between muscle, peritumoral inflammation and fibrosis (23, 24). The significance of dynamic contrast enhanced MRI(DCE-MRI) in assessing tumor aggressiveness depends on the neoangiogenesis of the tumor, which is an important factor in tumor growth; the more neovascularization there is, the higher the tumor stage and grade (25). In studies on dynamic enhancement sequences, the tumor, bladder mucosa and submucosa show early enhancement, but the bladder wall muscle maintains its low signal and delays enhancement. As early as 2000, Hayashi et al. observed that image signs of submucosal linear enhancement (SLE) at the base of the tumor were frequently seen on DCE images of NMIBC patients (26). This discovery is unquestionably a watershed moment in imaging-based BC staging and muscle-invasive status (MIS) diagnosis. Takeuchi et al (27) followed up by reporting an important feature found in most NMIBC on DWI, the tumor stalk, which improved the accuracy and robustness of imaging-based BC staging and MIS diagnosis. The accuracy of staging based on tumor stalk was 91.3% in Wang et al. study, while the accuracy of SLE staging was 91.3% (23). Panebianco et al (28) proposed Vesical Imaging-Reporting and Data System (VI-RADS) to quantify these signs on Multi-Parametric Magnetic Resonance Imaging (mpMRI) and to standardize the diagnostic procedure for image-based MIS prediction based on these features. This scoring system has effective diagnostic performance. In the Ueno et al. study, for example, the combined area under the curve (AUC) of five radiologists diagnosing MIBC was as high as 0.90 (29). Another prospective study also demonstrated the high diagnostic reliability of the VI-RADS score (AUC value of 0.94), especially for scores 1-2 and 3-5 (sensitivity 91.9%, 95%; specificity 91.1%, 95%) (30). The VI-RADS scoring method relies on expert visual perception judgment, yet it is still semi-qualitative. As a result, research into the objective and accurate radiomic detection of bladder cancer muscle invasion is required.

4.2 Radiomics

Radiomics is a relatively young concept, and Prof. Lambin originally described it in 2012 (31). Radiomics refers to the high-throughput extraction of image features from the region of interest (ROI) of radiological imaging techniques (CT, MR, but also PET, etc.) for automated analysis, using machine and deep learning techniques to extract critical information for accurate quantitative assessment of lesions, and ultimately for aiding in the diagnosis, classification, or grading of diseases. Radiomics inherits the technological benefits of reproducible, non-invasive radiological imaging over biopsy, making patient status monitoring and prognosis safer and more reliable.

Radiomics techniques can be classified into two groups: those using manual radiomics features and those using deep learning radiomics (32, 33) Traditional manual radiomics has the following four main processing tasks: image acquisition and preprocessing; image segmentation; feature extraction and quantification; and model building. The difference is that segmentation is not a necessity in the automated radiomics pipeline (33).. Radiomics has been increasingly studied in medical field for lung cancer, breast cancer, glioma, prostate cancer and other disorders (34–37). One of the current topics in bladder cancer research is the radiomics prediction of MIBC.

The pertinent radiomics literature is described below in terms of modality selection, volumes of interest (VOIs) segmentation, feature selection, model construction, and integration of clinical features, respectively.

4.2.1 Input modality

It mainly based on enhanced CT, MRI, with MRI accounting for (8/12) of the included literature. Since CT is weaker than MRI in discriminating soft tissues and the borders and bases of lesions are rarely distinguishable in discriminating MIS (38), there is a greater preference for MRI, mainly around T2WI, DWI and ADC and DCE sequences. In 2017, Garapati (11) and Xu et al. (12)established a precedent for using radiomics to predict MIS using CT and MRI, respectively, and inspired readers to combine additional MRI sequences to improve the possibility of differentiation task performance. As a result, extensive research on the precise differentiation of NMIBC and MIBC using radiomic methods with multi-parametric MRI images started to be conducted. Xu et al. obtained mean accuracies of 79.63%, 81.37%, and 91.22% for T2WI, DWI, and the combined of both sequences, with AUCs of 0.8828, 0.8884, and 0.9756, respectively (14). The superiority of DWI sequences over T2WI sequences in reflecting heterogeneous differences between NMIBC and MIBC (14, 16) has been repeatedly demonstrated. This might be because muscle-infiltrating tumors have a propensity to impede water molecule diffusion by shrinking extracellular space (39–41), which is better captured by DWI and the related ADC maps. And multi-sequence MRI was more helpful to predict the muscle invasion condition of BC preoperatively compared with single sequence T2WI and DWI, which was consistent with previous knowledge.

4.2.2 Volumes of interest segmentation

The three basic methods of delineating the area of interest are manual, semi-automated, and automatic. Even with computerized techniques, radiologists still need to examine and manually adjust them to assure the correctness of ROI descriptions because the majority of them are still primarily manual, which takes time and is tiresome. Initially, academics mostly concentrated on the overall tumor volume. As research developed, it was generally

acknowledged that the information in the region around the tumor also held a lot of relevant information. The body of literature suggests that the determination of muscle invasiveness is related to bladder tumors as well as the tumor's base (15) and adjacent bladder tissue (13). In addition, most of the relevant experiments have been conducted so far at the 3D level. Compared to 2D system analysis, 3D has higher precision and AUC (95.24% and 0.9864 *vs.* 92.86% and 0.9705) (18) which reflects the importance of 3D processing as it provides a comprehensive BC assessment with full descriptive information and details.

4.2.3 Feature extraction and quantization

Currently, there are mainly shape and intensity features based on histogram, texture features including gray level co-occurrence matrix (GLCM), gray Level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), gray level dependence matrix (GLDM), neighborhood gray-tone difference matrix (NGTDM), and higher-order feature wavelet features. The global, local and regional distribution features of image grayscale can be comprehensively described. Although there are a large number of features available for analysis, redundancy of features can seriously affect prediction performance. So feature selection is essential for developing optimal prediction models. Combined with other advanced selection strategies for statistical analysis, such as support vector machine (SVM)-based recursive feature elimination (SVM-RFE), the least absolute shrinkage and selection operator

(LASSO), max-relevance and min-redundancy(mRMR), these methods are widely used to reduce the impact of feature redundancy, and other methods such as Boruta are also used. After feature selection, Xu et al. found that the run length matrix (RLM) features accounted for a greater proportion of 13/19 in the optimal subset (14), better reflecting the regional heterogeneity differences between NMIBC and MIBC. The Co-occurrence matrices(CM), RLM and GLSZM features were found to be favorable feature classes for predicting BCa muscle invasion condition by Wang et al. (16).

4.2.4 Model construction

Different machine learning classifiers can be employed with the chosen features to create predictive models. Classifiers that are typically used include LASSO, SVM, random forest (RF), logistic regression, etc. Convolutional neural networks (CNN) are the most commonly used artificial neural networks for deep learning. SVM-RFE was the most commonly used machine learning method (7/12), among all the methods used for the classification task. Table 2 demonstrates how different models' prediction efficacy varies. NN, SVM, and RF classifier diagnostic performance were tested by Hammouda et al. in descending order (18). Garapati et al. observed that the AUC for morphological and texture features was roughly 0.90 (11); for various other mri-based radiomics models, the AUC ranged from 0.87 to 0.98 (14–17). However, all of the preceding experiments have the disadvantage of lacking independent external validation, so the true validity of the diagnostic performance of these models must be confirmed further. In contrast, so far, the prediction model developed by Zhang et al. is the only experiment with external validation results. But the AUC (0.791-0.936) of the study by Zhang et al. was slightly lower (19). This may be the risk of misclassification of some models influenced by tumor size, which may lead to a decrease in the diagnostic performance of the model, and therefore tumor size is one of the critical features to determine the muscle invasion condition of BC.

4.2.5 Integration of other clinical factors

It has become a trend to include clinical risk factors in the prediction model in order to better predict MIS and improve clinical diagnostic performance and application value. These include tumor size (15), tumor stalk (16), proteinuria and multiple sclerosis (21), as well as VI-RADS (20) and TURBT (14).The radiomic model incorporating clinical factors performed significantly better than the conventional MRI examination and simply radiomic model in terms of calibration and discrimination. Radiomic-clinical nomogram can be used as a reliable and non-invasive adjunct to differentiate MIBC from NMIBC preoperatively (15).

4.2.6 Method for validating results

83.3 percent (10/12) of the retrieved literature were single-center studies (11–15, 17, 18, 20–22), and the internal validation method was primarily used for model validation. Only two paper performing external validation of the results (16, 19). Because of the lack of externally validated results, the reliability of the remaining articles' results in terms of diagnostic efficacy is questionable. The sensitivity, specificity, and AUC of the internal validation cohort in Zhang's prediction model were 0.733, 0.810, and 0.861, respectively, while those of the external validation cohort were 0.710, 0.773, and 0.791, respectively (19).

5 Future and prospects

Of these 12 studies, all were retrospective, subject to selection bias and prone to data loss. Because the sample size was insufficient, cross-validation was essentially required to make up for it. Additionally, only two of the results were externally validated using radiomics models, with the rest being single-center, internally validated results that were not convincing. The current radiomics models are mainly based

TABLE 2 Radiomic characteristics of studies included in the systematic review.

Author	Segmentation method	Radiomic feature categories	Machine-learning method for feature selection	Number of selected features	Model	AUC of radiomic model with the best performance		clinical factor	AUC of radiomic-clinical model	
						Training set	Validation set		Training set	Validation set
Xu	Semi-automatic segmentation(3D)	Signal intensity histogram-based features and 3D ND-Haralick texture features based intensity and its high-order derivative maps	SVM-RFE","SMOTE	13	SVM-RFE	0.861	NA	NA	NA	NA
Garapati	Automatic segmentation(3D)	First-order statistics, shape, contrast, GLRLM,	Stepwise feature selection	3 subsets of radiomic features	LDA, NN, SVM, RF	0.97	NA	NA	NA	NA
Tong	Manual segmentation(3D)	LBP、GLCM	An optimal biomarker approach	9	SVM	Patient level:0.806, radial sector level:0.813	NA	NA	NA	NA
Xu	Manual segmentation(2D)	Histogram, CM , RLM,	SVM-RFE","SMOTE,	19	SVM-RFE	0.9857	NA	NA	NA	NA
Zheng	Semi-automatic segmentation(3D)	first-order statistics,shape-based,GLCM, GLRLM,GLSZM,NGTDM,and GLDM	LASSO LR	23	LASSO	0.913; Optimism-corrected:0.912	0.874	Tumor size	0.922; optimism-corrected AUC of 0.921	0.876
Xu	Manual segmentation and automatic segmentation(3D)	First-order intensity features,high-order texture features,and shape ,GLCM,GLRLM, GLSZM and NGTDM	Boruta	21	RF, AR	.0.907	0.904	RandomForest model and TURBT	NA	NA
Wang	Manual segmentation and automatic segmentation(2D)	Histogram , CM, RLM, NGTDM and GLSZM	SVM-RFE	36	LR, LASSO	0.88	external validation cohort 0.813	Radscore and tumor stalk	0.924	0.877
Hammouda	Automatic segmentation(3D)	Histogram ,GLCM,GLRLM,and morphological features	NA	NA	NN(best)","RF, SVM	0.9864		NA	NA	NA
Zhang	Semi-automatic segmentation(3D)	NA	NA	NA	FGP-Net	development cohort:0.936, tuning cohort:0.891	internal validation cohort: 0.861,external validation cohort: 0.791	NA	NA	NA
Zheng	manual segmentation (3D)	shape and size-based features, image intensity, textural features and wavelet features	mRMR	40	Lasso(best)、SVM、RF	0.934	0.906	VI-RADS	0.97	0.943

(Continued)

10.3389/fonc.2022.990176

TABLE 2 Continued

Author	Segmentation method	Radiomic feature categories	Machine-learning method for feature selection	Number of selected features	Model	AUC of radiomic model with the best performance		clinical factor	AUC of radiomic-clinical model	
						Training set	Validation set		Training set	Validation set
Zhou	semi-automatic segmentation(3D)	GLDM,Shape2D,GLCM,Shape3D,First-order,GLRLM,GLSZM,and NGTDM	SVM-RFE	6	LR, Decision Tree, SVM(best), and Adaboost algorithm	0.898	0.702	Rad-score, albuminuria and metabolic syndrome	0.8457	

AR, all-relevant model; AUC, area under the curve; CM, Co-occurrence matrices; 3D, three dimensional; 2D, two dimensional; FGP-Net, Filter-guided Pyramid Network; GLCM, grey-level co-occurrence matrix; GLDM, gray level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; LASSO, least absolute shrinkage and selection operator; LBP, local binary pattern; LDA, linear discriminant analysis; LR, logistic regression; mRMR, min-redundancy; NA, not available; ND, nondirectional Haralic textural features; NN, neural network; NGTDM, neighborhood gray-tone difference matrix; RF, random forest model; RFE, recursive feature elimination; RLM, run length matrix; SMOTE, synthetic minority oversampling technique; SVM, support vector machine classifier; TURBT, transurethral resection of bladder tumor.

on single-modality or dual-modality MRI, and there is no multi-modality study combining the three sequences of "T2WI, DWI and DCE", which needs to be further validated to improve the differentiation performance. Therefore, investigations should be planned in a more thorough and subtle manner for a variety of therapeutic applications to increase the reliability of the results. To completely understand the diagnostic usefulness of machine learning in predicting MIBC, more prospective multi-center and various machine trials will be required in the future. In addition, for future optimization of this new approach, more studies are needed to test the potential of optimizing predictive models by combining imaging biomarkers with other non-imaging biomarkers, such as urine and serum biomarkers. Although there have been significant advances in a number of studies, from fundamental tumor identification to precise staging and grading, recent research has also been gradually moving toward the prediction of treatment outcomes. The needs of the clinical market can no longer be met by illness diagnosis alone. After a bladder cancer diagnosis, increasing focus will be placed on how well machine learning predicts the response to treatment and prognosis outcome of the disease In the future.

Author contributions

FL and YL contributed to the conception of the study. XW contributed significantly to analysis and manuscript preparation. XH performed the data analyses and wrote the manuscript. XL and JD organized and drew tables. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by China International Medical Foundation (No. Z-2014-07-2101).

Conflict of interest

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.

Publisher's note

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660

2. Chavan S, Bray F, Lortet-Tieulent J, Goodman M, Jemal A. International variations in bladder cancer incidence and mortality. *Eur Urol* (2014) 66(1):59–73. doi: 10.1016/j.eururo.2013.10.001

3. Verma S, Rajesh A, Prasad SR, Gaitonde K, Lall CG, Mouraviev V, et al. Urinary bladder cancer: Role of MR imaging. *Radiographics* (2012) 32(2):371–87. doi: 10.1148/rg.322115125

4. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* (2006) 49(3):466–5; discussion 475-7. doi: 10.1016/j.eururo.2005.12.031

5. Sherif A, Jonsson MN, Wiklund NP. Treatment of muscle-invasive bladder cancer. *Expert Rev Anticancer Ther* (2007) 7(9):1279–83. doi: 10.1586/14737140.7.9.1279

6. Josephson D, Pasin E, Stein JP. Superficial bladder cancer: Part 2. management. *Expert Rev Anticancer Ther* (2007) 7(4):567–81. doi: 10.1586/14737140.7.4.567

7. Turker P, Bostrom PJ, Wroclawski ML, van Rhijn B, Kortekangas H, Kuk C, et al. Upstaging of urothelial cancer at the time of radical cystectomy: Factors associated with upstaging and its effect on outcome. *BJU Int* (2012) 110(6):804–11. doi: 10.1111/j.1464-410X.2012.10939.x

8. Ferro M, de Cobelli O, Musi G, Del Giudice F, Carrieri G, Busetto GM, et al. Radiomics in prostate cancer: An up-to-date review. *Ther Adv Urol* (2022) 14:17562872221109020. doi: 10.1177/17562872221109020

9. Tătaru OS, Vartolomei MD, Rassweiler JJ, Virgil O, Lucarelli G, Porpiglia F, et al. Artificial intelligence and machine learning in prostate cancer patient management-current trends and future perspectives. *Diagnostics (Basel)* (2021) 11(2):354. doi: 10.3390/diagnostics11020354

10. Ge L, Chen Y, Yan C, Zhao P, Zhang P, A R, et al. Study progress of radiomics with machine learning for precision medicine in bladder cancer management. *Front Oncol* (2019) 9:1296. doi: 10.3389/fonc.2019.01296

11. Garapati SS, Hadjiiski L, Cha KH, Chan HP, Caoili EM, Cohan RH, et al. Urinary bladder cancer staging in CT urography using machine learning. *Med Phys* (2017) 44(11):5814–23. doi: 10.1002/mp.12510

12. Xu X, Liu Y, Zhang X, Tian Q, Wu Y, Zhang G, et al. Preoperative prediction of muscular invasiveness of bladder cancer with radiomic features on conventional MRI and its high-order derivative maps. *Abdom Radiol (NY)* (2017) 42(7):1896–905. doi: 10.1007/s00261-017-1079-6

13. Tong Y, Udupa JK, Wang C, Chen J, Venigalla S, Guzzo TJ, et al. Radiomics-guided therapy for bladder cancer: Using an optimal biomarker approach to determine extent of bladder cancer invasion from t2-weighted magnetic resonance images. *Adv Radiat Oncol* (2018) 3(3):331–8. doi: 10.1016/j.adro.2018.04.011

14. Xu X, Zhang X, Tian Q, Wang H, Cui LB, Li S, et al. Quantitative identification of nonmuscle-invasive and muscle-invasive bladder carcinomas: A multiparametric MRI radiomics analysis. *J Magn Reson Imaging* (2019) 49(5):1489–98. doi: 10.1002/jmri.26327

15. Zheng J, Kong J, Wu S, Li Y, Cai J, Yu H, et al. Development of a noninvasive tool to preoperatively evaluate the muscular invasiveness of bladder cancer using a radiomics approach. *Cancer* (2019) 125(24):4388–98. doi: 10.1002/cncr.32490

16. Wang H, Xu X, Zhang X, Liu Y, Ouyang L, Du P, et al. Elaboration of a multisequence MRI-based radiomics signature for the preoperative prediction of the muscle-invasive status of bladder cancer: a double-center study. *Eur Radiol* (2020) 30(9):4816–27. doi: 10.1007/s00330-020-06796-8

17. Xu S, Yao Q, Liu G, Jin D, Chen H, Xu J, et al. Combining DWI radiomics features with transurethral resection promotes the differentiation between muscle-invasive bladder cancer and non-muscle-invasive bladder cancer. *Eur Radiol* (2020) 30(3):1804–12. doi: 10.1007/s00330-019-06484-2

18. Hammouda K, Khalifa F, Soliman A, Ghazal M, El-Ghar MA, Badawy MA, et al. A multiparametric MRI-based CAD system for accurate diagnosis of bladder cancer staging. *Comput Med Imaging Graph* (2021) 90:101911. doi: 10.1016/j.compmedimag.2021.101911

19. Zhang G, Wu Z, Xu L, Zhang X, Zhang D, Mao L, et al. Deep learning on enhanced CT images can predict the muscular invasiveness of bladder cancer. *Front Oncol* (2021) 11:654685. doi: 10.3389/fonc.2021.654685

20. Zheng Z, Xu F, Gu Z, Yan Y, Xu T, Liu S, et al. Combining multiparametric MRI radiomics signature with the vesical imaging-reporting and data system (VI-RADS) score to preoperatively differentiate muscle invasion of bladder cancer. *Front Oncol* (2021) 11:619893. doi: 10.3389/fonc.2021.619893

21. Zhou Q, Zhang Z, Ang X, Zhang H, Ouyang J. A nomogram combined with radiomics features, albuminuria, and metabolic syndrome to predict the risk of myometrial invasion of bladder cancer. *Transl Cancer Res* (2021) 10(7):3177–91. doi: 10.21037/tcr-21-426

22. Cui Y, Sun Z, Liu X, Zhang X, Wang X. CT-based radiomics for the preoperative prediction of the muscle-invasive status of bladder cancer and comparison to radiologists' assessment. *Clin Radiol* (2022) 77(6):e473–82. doi: 10.1016/j.crad.2022.02.019

23. Wang HJ, Pui MH, Guo Y, Yang D, Pan BT, Zhou XH, et al. Diffusion-weighted MRI in bladder carcinoma: the differentiation between tumor recurrence and benign changes after resection. *Abdom Imaging* (2014) 39(1):135–41. doi: 10.1007/s00261-013-0038-0

24. Wang HJ, Pui MH, Guan J, Li SR, Lin JH, Pan B, et al. Comparison of early submucosal enhancement and tumor stalk in staging bladder urothelial carcinoma. *AJR Am J Roentgenol* (2016) 207(4):797–803. doi: 10.2214/AJR.16.16283

25. Abouelkheir RT, Abdelhamid A, Abou El-Ghar M, El-Diasty T. Imaging of bladder cancer: Standard applications and future trends. *Medicina (Kaunas)* (2021) 57(3):220. doi: 10.3390/medicina57030220

26. Hayashi N, Tochigi H, Shiraishi T, Takeda K, Kawamura J. A new staging criterion for bladder carcinoma using gadolinium-enhanced magnetic resonance imaging with an endorectal surface coil: A comparison with ultrasonography. *BJU Int* (2000) 85(1):32–6. doi: 10.1046/j.1464-410x.2000.00358.x

27. Takeuchi M, Sasaki S, Ito M, Okada S, Takahashi S, Kawai T, et al. Urinary bladder cancer: Diffusion-weighted MR imaging–accuracy for diagnosing T stage and estimating histologic grade. *Radiology* (2009) 251(1):112–21. doi: 10.1148/radiol.2511080873

28. Panebianco V, Narumi Y, Altun E, Bochner BH, Efstathiou JA, Hafeez S, et al. Multiparametric magnetic resonance imaging for bladder cancer: Development of VI-RADS (Vesical imaging-reporting and data system). *Eur Urol* (2018) 74(3):294–306. doi: 10.1016/j.eururo.2018.04.029

29. Ueno Y, Takeuchi M, Tamada T, Sofue K, Takahashi S, Kamishima Y, et al. Diagnostic accuracy and interobserver agreement for the vesical imaging-reporting and data system for muscle-invasive bladder cancer: A multireader validation study. *Eur Urol* (2019) 76(1):54–6. doi: 10.1016/j.eururo.2019.03.012

30. Del Giudice F, Barchetti G, De Berardinis E, Pecoraro M, Salvo V, Simone G, et al. Prospective assessment of vesical imaging reporting and data system (VI-RADS) and its clinical impact on the management of high-risk non-muscle-invasive bladder cancer patients candidate for repeated transurethral resection. *Eur Urol* (2020) 77(1):101–9. doi: 10.1016/j.eururo.2019.09.029

31. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48(4):441–6. doi: 10.1016/j.ejca.2011.11.036

32. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol* (2022) 19(2):132–46. doi: 10.1038/s41571-021-00560-7

33. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol* (2019) 16(7):391–403. doi: 10.1038/s41585-019-0193-3

34. Smith CP, Czarniecki M, Mehralivand S, Stoyanova R, Choyke PL, Harmon S, et al. Radiomics and radiogenomics of prostate cancer. *Abdom Radiol (NY)* (2019) 44(6):2021–9. doi: 10.1007/s00261-018-1660-7

35. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. *Breast* (2020) 49:74–80. doi: 10.1016/j.breast.2019.10.018

36. Choi YS, Bae S, Chang JH, Kang SG, Kim SH, Kim J, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas *via* deep learning and radiomics. *Neuro Oncol* (2021) 23(2):304–13. doi: 10.1093/neuonc/noaa177

37. Kolinger GD, García DV, Kramer GM, Frings V, Zwezerijnen GJC, Smit EF, et al. Effects of tracer uptake time in non-small cell lung cancer (18)F-FDG PET radiomics. *J Nucl Med* (2022) 63(6):919–24. doi: 10.2967/jnumed.121.262660

38. Xu X, et al. Study progress of noninvasive imaging and radiomics for decoding the phenotypes and recurrence risk of bladder cancer. *Front Oncol* (2021) 11:704039. doi: 10.3389/fonc.2021.704039

39. Hafeez S, Huddart R. Advances in bladder cancer imaging. *BMC Med* (2013) 11:104. doi: 10.1186/1741-7015-11-104

40. Mannelli L, Nougaret S, Vargas HA, Do RK. Advances in diffusion-weighted imaging. *Radiol Clin North Am* (2015) 53(3):569–81. doi: 10.1016/j.rcl.2015.01.002

41. Zhang X, Xu X, Tian Q, Li B, Wu Y, Yang Z, et al. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. *J Magn Reson Imaging* (2017) 46(5):1281–8. doi: 10.1002/jmri.25669

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 18 August 2022
DOI 10.3389/fonc.2022.985940

OPEN ACCESS

EDITED BY
Xiongbing Zu,
Xiangya Hospital, Central South University, China

REVIEWED BY
Bin Xu,
Shanghai Jiao Tong University, China
Boyang Wu,
Washington State University, United States
Zilian Cui,
Shandong Provincial Hospital, China

*CORRESPONDENCE
Jiwen Cheng
chengjiwen1977@foxmail.com

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 04 July 2022
ACCEPTED 28 July 2022
PUBLISHED 18 August 2022

CITATION
Deng X, Li T, Mo L, Wang F, Ji J, He X, Mohamud BH, Pradhan S and Cheng J (2022) Machine learning model for the prediction of prostate cancer in patients with low prostate-specific antigen levels: A multicenter retrospective analysis.
Front. Oncol. 12:985940.
doi: 10.3389/fonc.2022.985940

Machine learning model for the prediction of prostate cancer in patients with low prostate-specific antigen levels: A multicenter retrospective analysis

Xiaobin Deng[1,2], Tianyu Li[1,2,3], Linjian Mo[1,2,3], Fubo Wang[1,2,3], Jin Ji[4], Xing He[4], Bashir Hussein Mohamud[1,2], Swadhin Pradhan[1,2] and Jiwen Cheng[1,2,3]*

[1]Department of Urology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [2]Institute of Urology and Nephrology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [3]Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, China, [4]Department of Urology, Changhai Hospital, Naval Medical University, Shanghai, China

Objective: The aim of this study was to develop a predictive model to improve the accuracy of prostate cancer (PCa) detection in patients with prostate specific antigen (PSA) levels ≤20 ng/mL at the initial puncture biopsy.

Methods: A total of 146 patients (46 with Pca, 31.5%) with PSA ≤20 ng/mL who had undergone transrectal ultrasound-guided 12+X prostate puncture biopsy with clear pathological results at the First Affiliated Hospital of Guangxi Medical University (November 2015 to December 2021) were retrospectively evaluated. The validation group was 116 patients drawn from Changhai Hospital(52 with Pca, 44.8%). Age, body mass index (BMI), serum PSA, PSA-derived indices, several peripheral blood biomarkers, and ultrasound findings were considered as predictive factors and were analyzed by logistic regression. Significant predictors ($P < 0.05$) were included in five machine learning algorithm models. The performance of the models was evaluated by receiver operating characteristic curves. Decision curve analysis (DCA) was performed to estimate the clinical utility of the models. Ten-fold cross-validation was applied in the training process.

Results: Prostate-specific antigen density, alanine transaminase-to-aspartate transaminase ratio, BMI, and urine red blood cell levels were identified as independent predictors for the differential diagnosis of PCa according to multivariate logistic regression analysis. The RandomForest model exhibited the best predictive performance and had the highest net benefit when compared with the other algorithms, with an area under the curve of 0.871. In addition, DCA had the highest net benefit across the whole range of cut-off points examined.

Conclusion: The RandomForest-based model generated showed good prediction ability for the risk of PCa. Thus, this model could help urologists in the treatment decision-making process.

KEYWORDS

prostate cancer, prostate-specific antigen, diagnosis, machine learning, predictive model, real-world study

Introduction

Prostate cancer (PCa) remains the most common malignancy in men. According to the latest cancer statistics published in 2022, PCa accounts for 27% of newly diagnosed malignancies in males, and is also the second leading cause of death among men with cancer (1). Serum total prostate-specific antigen (tPSA) is a specific tumor biomarker of PCa in the clinical setting. It has high tissue specificity, but is also associated with missed diagnoses and misdiagnoses (2, 3). A variety of benign diseases of the prostate, including benign prostatic hyperplasia and prostatitis, can lead to elevated serum tPSA levels (4, 5). In particular, tPSA levels in the range of 4 to 20 ng/mL are associated with a PCa incidence of less than 25%. In addition, patients with serum tPSA levels ≤4 ng/mL still carry the risk of PCa, and the detection rate in these patients may reach up to 20% (6, 7). Prostate puncture biopsy is currently the standard modality for diagnosing PCa, but as an invasive procedure, it carries a risk of infection. Moreover, the PCa detection rates on initial prostate puncture biopsies range from 23% to 42%. These limitations have greatly restricted its clinical use (8–11).

Given the limitations of the currently used diagnostic methods, a large number of studies are currently devoted to identifying new predictors of PCa. For example, PSA-derived indices, such as free-to-total PSA (F/T) values and prostate-specific antigen density (PSAD), have been found to exhibit greater diagnostic accuracy than PSA alone (12–14). In addition, several inflammatory and neurotrophic markers, including neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, monocyte-to-lymphocyte ratio (15, 16), gamma-glutamyl transpeptidase-to-lymphocyte count ratio (17, 18), red cell distribution width-to-platelet ratio (19, 20), and alanine transaminase-to-aspartate transaminase ratio (21, 22), have previously been shown to have predictive value not only as inflammatory markers but also as indicators for the diagnosis and prognosis of malignancies. To date, studies focusing on the role of these composite indicators in the differential diagnosis of PCa are limited.

Machine learning (ML), as an important branch of artificial intelligence, can continuously optimize the performance of predictive or diagnostic models by learning and analyzing data, and can handle non-linear relationships better than traditional statistical scores. As a result, ML-based models have great potential for the diagnosis and prognosis of diseases (23–25). Therefore, our goal was to develop a new decision-support ML model based on real-world data for diagnosing PCa in patients with PSA levels ≤20 ng/mL.

Materials and methods

Ethics statement

This study was approved by the institutional review board of the First Affiliated Hospital of Guangxi Medical University. Written informed consent was obtained from all patients for the storage of their information for the purpose of research. All the research procedures were conducted in accordance with the Declaration of Helsinki.

Data collection

Data from patients with PCa or benign prostatic hyperplasia who underwent systematic prostate puncture biopsy at our hospital between November 2015 and December 2021 were collected and retrospectively analyzed. We included adult patients with tPSA levels ≤20 ng/ml who underwent transrectal ultrasound (TRUS)-guided prostate puncture biopsy for at least systemic 12 cores with clear pathological results. The exclusion criteria were (1) a history of prostate cancer, prostate surgery, or 5-alpha-reductase inhibitor/drug for the treatment of endocrine dyscrasia in prostate cancer; (2) a diagnosis of prostatitis; (3) digital rectal examination (DRE), transrectal ultrasonography, or cystoscopy within two weeks before PSA detection (as these examinations may affect serum PSA levels); and (4) missing hematological data prior to puncture biopsy. Serum PSA concentrations (tPSA and fPSA) were measured before DRE and TRUS by enzyme-linked immunoassay. Prostate volume (PV) was calculated using the following formula:

PV = 0.52 × anterior/posterior diameter (cm) × left/right diameter (cm) × upper/lower diameter (cm)

TRUS was performed by experienced ultrasonologists.

Statistical analysis

Continuous variables were converted into categorical variables. The optimum cutoff values obtained from ROC curve analysis were determined by maximizing the Youden index. Logistic regression analysis was applied to calculate the odds ratio (OR) with 95% confidence interval (CI). $P < 0.05$ was considered to indicate statistical significance. We used five different ML algorithms to analyze our data: logistic regression (LR), XGBoost (XGB), RandomForest (RF), multilayer perceptron (MLP), and k-nearest neighbor (kNN). After training, the model with the highest average AUC was chosen as the best algorithm. Furthermore, the ML-based model was tuned to avoid overfitting, and the accuracy of the algorithm was tested using the ten-fold cross-validation method. All variables were tested for Spearman correlations, and the results are presented as a heat map.

Results

Demographic features

A total of 146 eligible patients were included in this study. The optimal cut-off value of tPSA was 8.47 ng/mL, and the optimal cut-off value of BMI was 23.6 kg/m^2. The detailed clinical characteristics of all the patients are presented in Table 1. Among the evaluated clinical characteristics, PSA, F/T, BMI, alanine transaminase-to-aspartate transaminase ratio (LSR), red cell volume distribution width (RDW), alkaline phosphatase (ALP), and urine RBC level were correlated with the risk of PCa. Based on the correlation heat map (Figure 1), eight highly correlated features were chosen as predictors. Weight is inextricably linked to BMI and, to a lesser degree, height. Therefore, we used BMI instead of height or weight, since it is a better indicator of obesity. The external validation cohort was screened based on inclusion and exclusion criteria consistent with the training cohort. Most of the externally validated variables did not differ statistically from the training cohort.

Univariate and multivariate logistic regression analyses

According to univariate logistic regression analysis (Table 2), tPSA, PSAD, F/T, BMI, LSR, RDW, and blood glucose level were significant predictors of the occurrence of PCa in the overall population ($P > 0.05$). PSA and PSAD are significantly correlated, and the univariate analysis indicated that PSAD was more statistically significant as a predictor than PSA. Therefore, we only included PSAD in the multivariate analysis. The significant characteristics identified from the univariate analysis above were included in multivariate logistic regression analysis (Table 3). The odds ratios (ORs) calculated indicated the relative risk of PCa. The results showed that PSAD (OR = 11.539, 95% CI = 4.388–33.993), LSR (OR = 0.189, 95% CI = 0.059–0.561), BMI (OR = 2.638, 95% CI = 1.067–6.871), and urine RBC level (OR = 0.136, 95% CI = 0.018–0.620) were independent predictors of PCa. In addition, ALP (OR = 6.00, 95% CI = 1.052–34.212) was also identified as a significant independent predictor ($P = 0.044$) in univariate logistic regression analysis, but it was not included in further analysis as there were too many missing values (n = 31), Supplementary Tables 2, 3.

Performance of ML algorithms

To compare the predictive performance of the six ML algorithm models, ten-fold cross validation and decision curve analysis was applied (Figure 2). As shown in the figure, the RF model exhibited the best performance in the differential diagnosis of PCa, with an average AUC of 0.871 (95% CI = 0.808–0.933). The ML algorithm-based models outperformed PSA and its derivatives F/T and PSAD individually by a significant margin (AUC of PSA = 0.589, AUC of F/T = 0.599, AUC of PSAD = 0.737). Therefore, the RF model was finally regarded as the preferred prediction model. In the external validation group (Figure 3), RF (AUC = 0.780, 95% CI = 0.691–0.869), LR (AUC = 0.781, 95% CI = 0.692–0.871) and XGB (AUC = 0.780; 95% CI = 0.692–0.868) showed good AUC values of 0.780. Based on the findings for the training cohort and the external validation cohort together, we finally choose the RF algorithm model as the best model and used it for further analysis.

Relative importance of the analyzed variables

The importance of the included features based on the RF algorithm differed from each other, and PSAD was identified as the most important variable. They were arranged as follows in descending order of importance: PSAD, LSR, urine RBC level, and BMI (Figure 4).

Discussion

One of the main topics of research on urological prostate cancer is the improvement of prediction accuracy before

TABLE 1 Clinical characteristics of patients in the training cohort.

Variables, n (%)	Level	Total	BPH	PCa	P-value
PSA;ng/mL	<8.47	59 (40.411)	46 (46.000)	13 (28.261)	0.042
	≥8.47	87 (59.589)	54 (54.000)	33 (71.739)	
fPSA;ng/mL	<1.89	98 (67.123)	63 (63.000)	35 (76.087)	0.118
	≥1.89	48 (32.877)	37 (37.000)	11 (23.913)	
F/T	<0.103	31 (21.233)	15 (15.000)	16 (34.783)	0.007
	≥0.103	115 (78.767)	85 (85.000)	30 (65.217)	
PV;mL	<38.1	52 (35.616)	20 (20.000)	32 (69.565)	<0.001
	≥38.1	94 (64.384)	80 (80.000)	14 (30.435)	
PSAD	<0.24	95 (65.068)	80 (80.000)	15 (32.609)	<0.001
	≥0.24	51 (34.932)	20 (20.000)	31 (67.391)	
Age;years	<73	118 (80.822)	85 (85.000)	33 (71.739)	0.059
	≥73	28 (19.178)	15 (15.000)	13 (28.261)	
BMI;kg/m²	<23.62	75 (51.370)	57 (57.000)	18 (39.130)	0.045
	≥23.624	71 (48.630)	43 (43.000)	28 (60.870)	
NLR	<1.46	29 (19.863)	16 (16.000)	13 (28.261)	0.085
	≥1.46	117 (80.137)	84 (84.000)	33 (71.739)	
PLR	<131.01	87 (59.589)	62 (62.000)	25 (54.348)	0.381
	≥131.01	59 (40.411)	38 (38.000)	21 (45.652)	
MLR	<0.336	84 (57.534)	55 (55.000)	29 (63.043)	0.361
	≥0.336	62 (42.466)	45 (45.000)	17 (36.957)	
GLR	<13.21	60 (41.096)	37 (37.000)	23 (50.000)	0.138
	≥13.21	86 (58.904)	63 (63.000)	23 (50.000)	
LSR	<0.684	32 (21.918)	16 (16.000)	16 (34.783)	0.011
	≥0.684	114 (78.082)	84 (84.000)	30 (65.217)	
RPR	<0.00063	66 (45.205)	40 (40.000)	26 (56.522)	0.062
	≥0.00063	80 (54.795)	60 (60.000)	20 (43.478)	
*WBC;*10⁹ /L*	<4.89	25 (17.123)	13 (13.000)	12 (26.087)	0.051
	≥4.89	121 (82.877)	87 (87.000)	34 (73.913)	
Hb;g/dl	<146.2	120 (82.192)	78 (78.000)	42 (91.304)	0.051
	≥146.2	26 (17.808)	22 (22.000)	4 (8.696)	
RDW; %	<0.15	125 (85.616)	81 (81.000)	44 (95.652)	0.019
	≥0.15	21 (14.384)	19 (19.000)	2 (4.348)	
*Plt;*10⁹ /L*	<207	71 (48.630)	52 (52.000)	19 (41.304)	0.23
	≥207	75 (51.370)	48 (48.000)	27 (58.696)	
*Neutrophil count;*10⁹ /L*	<4.14	84 (57.534)	54 (54.000)	30 (65.217)	0.203
	≥4.14	62 (42.466)	46 (46.000)	16 (34.783)	
*Lymphocyte count;*10⁹ /L*	<1.5	38 (26.027)	23 (23.000)	15 (32.609)	0.219
	≥1.5	108 (73.973)	77 (77.000)	31 (67.391)	
*Monocyte count;*10⁹ /L*	<0.44	32 (21.918)	19 (19.000)	13 (28.261)	0.209
	≥0.44	114 (78.082)	81 (81.000)	33 (71.739)	
*Eosinophil count;*10⁹ /L*	<0.17	69 (47.260)	42 (42.000)	27 (58.696)	0.061
	≥0.17	77 (52.740)	58 (58.000)	19 (41.304)	
Blood glucose;mmol/L	<5	62 (42.466)	37 (37.000)	25 (54.348)	0.049
	≥5	84 (57.534)	63 (63.000)	21 (45.652)	

(Continued)

TABLE 1 Continued

Variables, n (%)	Level	Total	BPH	PCa	P-value
γ-glutamyl transpeptidase;U	<33.2	98 (67.123)	72 (72.000)	26 (56.522)	0.064
	≥33.2	48 (32.877)	28 (28.000)	20 (43.478)	
Creatinine;μmol/L	<88	86 (58.904)	62 (62.000)	24 (52.174)	0.262
	≥88	60 (41.096)	38 (38.000)	22 (47.826)	
ALP;U	<90	21 (67.742)	18 (78.261)	3 (37.500)	0.034
	≥90	10 (32.258)	5 (21.739)	5 (62.500)	
Urine WBC	Negative	116 (79.452)	75 (75.000)	41 (89.130)	0.05
	Positive	30 (20.548)	25 (25.000)	5 (10.870)	
Urine RBC	Negative	121 (82.877)	77 (77.000)	44 (95.652)	0.005
	Positive	25 (17.123)	23 (23.000)	2 (4.348)	
Ultrasound hypoechoic nodules	Negative	46 (31.507)	30 (30.000)	16 (34.783)	0.563
	Positive	100 (68.493)	70 (70.000)	30 (65.217)	
Prostatic calculi	Negative	66 (45.205)	45 (45.000)	21 (45.652)	0.941
	Positive	80 (54.795)	55 (55.000)	25 (54.348)	

prostate puncture biopsy in order to reduce unnecessary patient pain without compromising on early intervention in patients with a confirmed diagnosis. The study variables included in this study were non-invasive, and data on these variables were readily available prior to biopsy. Therefore, obtaining the data for these variables did not involve any unduly demanding conditions or excessive medical overhead. Previous studies have mostly been limited to PSA and its derivatives, and there are not enough studies on the differential value of other inflammatory markers (26–28). Further, the currently used prediction nomogram based on the conventional algorithm also has room for further improvement.

The clinical application of ML algorithms may facilitate a paradigm shift in the medical field, as these algorithms are efficient, objective, and reproducible when it comes to large amounts of nonlinear data (24, 29–32). They also have the potential to improve the quality of early diagnosis, identify disease progression, and increase the likelihood of predicting patient-specific outcomes (25, 33, 34). These advantages can facilitate the sharing of information for decision-making

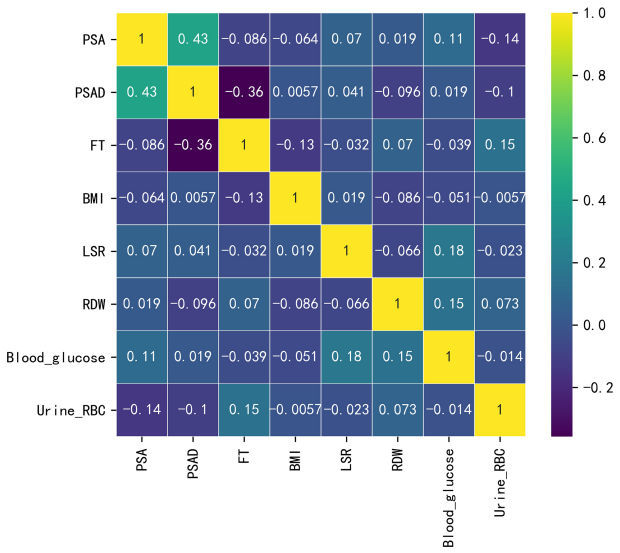


FIGURE 1
Heat map depicting the correlations between the examined variables.

TABLE 2 Univariate logistic regression in the differential diagnosis of prostate cancer in the whole data cohort.

Variables	OR	95%CI	P-value
PSA;ng/mL			
<8.47	1 (reference)		
≥8.47	2.162	[1.019,4.590]	0.045
fPSA;ng/mL			
<1.89	1 (reference)		
≥1.89	0.535	[0.243,1.179]	0.121
F/T			
<0.103	1 (reference)		
≥0.103	0.331	[0.146,0.750]	0.008
PV;mL			
<38.1	1 (reference)		
≥38.1	0.109	[0.049,0.243]	<0.001
PSAD			
<0.24	1 (reference)		
≥0.24	8.267	[3.761,18.169]	<0.001
Age;years			
<73	1 (reference)		
≥73	2.232	[0.959,5.194]	0.062
BMI;kg/m²			
<23.62	1 (reference)		
≥23.62	2.062	[1.011,4.204]	0.046
NLR			
<1.46	1 (reference)		
≥1.46	0.484	[0.210,1.115]	0.088
PLR			
<131.01	1 (reference)		
≥131.01	1.371	[0.676,2.779]	0.382
MLR			
<0.336	1 (reference)		
≥0.336	0.716	[0.350,1.467]	0.362
GLR			
<13.21	1 (reference)		
≥13.21	0.587	[0.290,1.190]	0.14
LSR			
<0.684	1 (reference)		
≥0.684	0.357	[0.159,0.802]	0.013
RPR			
<0.00063	1 (reference)		

(Continued)

TABLE 2 Continued

Variables	OR	95%CI	P-value
≥0.00063	0.513	[0.253,1.040]	0.064
*WBC;*10⁹ /L*			
<4.89	1 (reference)		
≥4.89	0.423	[0.176,1.020]	0.055
Hb;g/dl			
<146.2	1 (reference)		
≥146.2	0.338	[0.109,1.045]	0.06
RDW; %			
<0.15	1 (reference)		
≥0.15	0.194	[0.043,0.871]	0.032
*Plt;*10⁹ /L*			
<207	1 (reference)		
≥207	1.539	[0.760,3.119]	0.231
*Neutrophil count;*10⁹ /L*			
<4.14	1 (reference)		
≥4.14	0.626	[0.304,1.290]	0.204
*Lymphocyte count;*10⁹ /L*			
<1.5	1 (reference)		
≥1.5	0.617	[0.285,1.337]	0.221
*Monocyte count;*10⁹ /L*			
<0.44	1 (reference)		
≥0.44	0.595	[0.264,1.343]	0.212
*Eosinophil count;*10⁹ /L*			
<0.17	1 (reference)		
≥0.17	0.51	[0.251,1.035]	0.062
Blood glucose;mmol/L			
<5	1 (reference)		
≥5	0.493	[0.243,1.002]	0.05
γ-glutamyl transpeptidase;U			
<33.2	1 (reference)		
≥33.2	1.978	[0.955,4.097]	0.066
Creatinine;μmol/L			
<88	1 (reference)		
≥88	1.496	[0.739,3.028]	0.263
ALP;U			

(Continued)

10.3389/fonc.2022.985940

TABLE 2 Continued

Variables	OR	95%CI	P-value
<90	1 (reference)		
≥90	6	[1.052,34.212]	0.044
Urine WBC			
Negative	1 (reference)		
Positive	0.366	[0.130,1.028]	0.056
Urine RBC			
Negative	1 (reference)		
Positive	0.152	[0.034,0.676]	0.013
Ultrasound hypoechoic nodules			
Negative	1 (reference)		
Positive	0.804	[0.382,1.688]	0.564
Prostatic calculi			
Negative	1 (reference)		
Positive	0.974	[0.483,1.964]	0.941

between clinicians and patients and promote efficient planning and visualization of the use of healthcare services. In addition, the model can be actively retrained over time to continuously improve its own predictive accuracy.

The AUC value of our RF algorithm was 0.871 in the training cohort, and this value is significantly higher than the individual AUC values of PSA (AUC = 0.589), F/T PSA (AUC = 0.599), and PSAD (AUC = 0.737). The RF-based

TABLE 3 Multivariate logistic regression in the differential diagnosis of prostate cancer in the whole data cohort.

Variables	OR	95%CI	P-value
PSAD			
<0.24	1(reference)		
≥0.24	11.539	(4.388,33.993)	<0.001
F/T			
<0.103	1(reference)		
≥0.103	0.848	(0.294,2.515)	0.762
LSR			
<0.684	1(reference)		
≥0.684	0.189	(0.059,0.561)	0.004
BMI;kg/m²			
<23.62	1(reference)		
≥23.62	2.638	(1.067,6.871)	0.04
RDW; %			
<0.15	1(reference)		
≥0.15	0.259	(0.036,1.156)	0.111
Blood glucose;mmol/L			
<5	1(reference)		
≥5	0.501	(0.192,1.269)	0.148
Urine_RBC			
Negative	1(reference)		
Positive	0.136	(0.018,0.62)	0.022

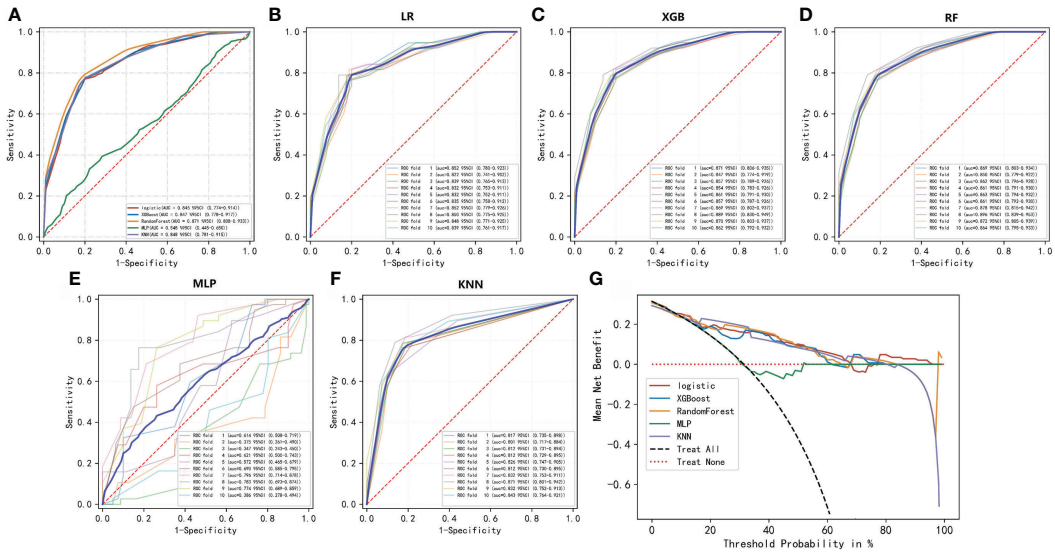


FIGURE 2
ROC and decision curve analyses of the five ML algorithms. **(A–F)** ROC curve analysis of a ten-fold cross-validation of five machine learning algorithms for predicting the risk of PCa in the training cohort. **(G)** Decision curve analysis demonstrating the net benefit associated with the use of the models for the prediction of upstaging.

model performed well in the external validation cohort, with an AUC value of 0.78, a sensitivity of 0.712, and a specificity of 0.85. Decision curve analysis was used to validate the efficacy and potential benefits of our novel model. This ML-based model can be used as a screening tool for prostate biopsy and has the potential to avoid missed diagnosis of PCa. Four independent predictors for PCa diagnosis were identified in our analysis: PSAD, BMI, LSR, and urine RBC level. Previous studies have suggested that obese patients have a higher risk of developing prostate cancer. This is probably because periprostatic fat is biologically active and can secrete factors that promote cancer growth. However, it is unclear whether

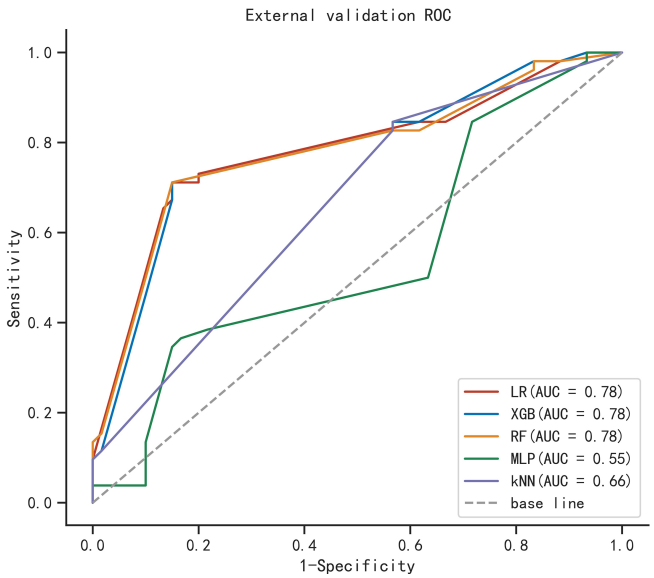


FIGURE 3
ROC curve analysis of five machine learning algorithms for predicting the risk of PCa in the external cohort.

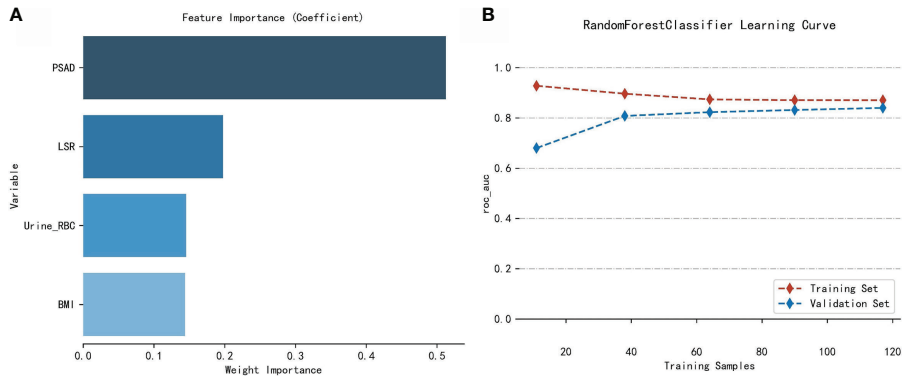


FIGURE 4
Analysis of the RF model. **(A)** Importance of the variables included in the RF model in the training cohort. **(B)** Learning curves of the RF model in the training cohort.

reversing obesity can mitigate the progression of prostate cancer (35–38). The presence of a visible hematuria is a common sign of prostatic disease. We believe that urine RBC level emerged as a predictive factor in this study because there was a high percentage of patients with benign lesions, and prostatic hyperplasia is associated with a lower incidence of urinary tract symptoms. Difficulty in urination can cause damage to the microvasculature of the urinary system, and this can manifest as urine occult blood. In contrast, PCa in its early clinical stage is often insidious, and most patients only seek treatment when elevated PSA is detected during routine physical examination. In previous studies, LSR has been applied in the evaluation of gestational diabetes (21), diagnosis of cirrhosis (22), and prognosis of different cancers (39, 40). The levels of alanine aminotransferase and aspartate aminotransferase may be affected by obesity (41, 42), and fluctuations in these two indicators may influence the diagnosis of PCa in a similar way that BMI influences PCa. Although ALP was not further analyzed in the current study due to missing data, ALP may still be a promising indicator for the diagnosis of prostate cancer. Further, it has been suggested that prostate cancer may exhibit overexpression of tumor-derived ALP, but this needs to be validated in further studies (43)

Our study has several limitations. First, the small sample size may affect the conclusions of the statistical analysis. Second, our study was a single-center retrospective analysis, so there is a possibility of a selection bias that may have affected the accuracy of our model. Future external validation is needed to assess the clinical application of our ML model by using data from other institutions. Finally, some meaningful indicators may not have been included in our analysis because of the absence of some hematological data, and this may have affected the efficacy of the model. Therefore, these findings need to be confirmed in future investigations on larger patient samples.

Conclusion

We established an efficient ML model for the differential diagnosis of PCa. Our model exhibited excellent predictive accuracy and practical clinical utility, and may help guide the decision-making process of the urologist, avoid unnecessary prostate puncture biopsy, and increase the detection rate of PCa.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

This study was approved by the institutional review board of the First Affiliated Hospital of Guangxi Medical University. Written informed consent was obtained from all patients for the storage of their information for the purpose of research. All the research procedures were conducted in accordance with the Declaration of Helsinki.

Author contributions

XD, TL, LM and FW: contributed to the conception and design. XD, JJ, XH, BM and SP: collected, analyzed the data, drew the figures and tables, and wrote the draft. XD and JC contributed to manuscript writing and revision. All authors approved the final manuscript.

Funding

This study was sponsored by Guangxi Science and Technology Base and Talent Project (Grant No. Guike AD20238090) and Guangxi Clinical Research Center for Urology and Nephrology (Grant No. Guike AD20297081).

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.985940/full#supplementary-material

SUPPLEMENTARY TABLE 1
Clinical characteristics of patients in the external validation cohort.

SUPPLEMENTARY TABLE 2
Raw data from the First Affiliated Hospital of Guangxi Medical University, Guangxi, China.

SUPPLEMENTARY TABLE 3
Validation database from Changhai Hospital, Naval Medical University, Shanghai, China.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin* (2022) 72:7–33. doi: 10.3322/caac.21708

2. Verma P, Malhotra G, Agrawal R, Sonavane S, Meshram V, Asopa RV. Evidence of prostate-specific membrane antigen expression in metastatic differentiated thyroid cancer using 68Ga-PSMA-HBED-CC PET/CT. *Clin Nucl Med* (2018) 43:e265–8. doi: 10.1097/RLU.0000000000002161

3. Aminsharifi A, Howard L, Wu Y, De Hoedt A, Bailey C, Freedland SJ, et al. Prostate specific antigen density as a predictor of clinically significant prostate cancer when the prostate specific antigen is in the diagnostic gray zone: Defining the optimum cutoff point stratified by race and body mass index. *J Urol.* (2018) 200:758–66. doi: 10.1016/j.juro.2018.05.016

4. Roehrborn CG, McConnell J, Bonilla J, Rosenblatt S, Hudson PB, Malek GH, et al. Serum prostate specific antigen is a strong predictor of future prostate growth in men with benign prostatic hyperplasia. PROSCAR long-term efficacy and safety study. *J Urol.* (2000) 163:13–20. doi: 10.1016/S0022-5347(05)67962-1

5. Mao Q, Zheng X, Jia X, Wang Y, Qin J, Yang K, et al. Relationships between total/free prostate-specific antigen and prostate volume in Chinese men with biopsy-proven benign prostatic hyperplasia. *Int Urol Nephrol.* (2009) 41:761–6. doi: 10.1007/s11255-009-9533-1

6. Ahyai SA, Graefen M, Steuber T, Haese A, Schlomm T, Walz J, et al. Contemporary prostate cancer prevalence among T1c biopsy-referred men with a prostate-specific antigen level < or = 4.0 ng per milliliter. *Eur Urol.* (2008) 53:750–7. doi: 10.1016/j.eururo.2007.10.017

7. Franlund M, Arnsrud GR, Carlsson SV, Lilja H, Mansson M, Stranne J, et al. Prostate cancer risk assessment in men with an initial P.S.a. below 3 ng/mL: Results from the goteborg randomized population-based prostate cancer screening trial. *Scand J Urol.* (2018) 52:256–62. doi: 10.1080/21681805.2018.1508166

8. Huang GL, Kang CH, Lee WC, Chiang PH. Comparisons of cancer detection rate and complications between transrectal and transperineal prostate biopsy approaches - a single center preliminary study. *BMC Urol.* (2019) 19:101. doi: 10.1186/s12894-019-0539-4

9. Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. part 1: Screening, diagnosis, and local treatment with curative intent. *Eur Urol.* (2017) 71:618–29. doi: 10.1016/j.eururo.2016.08.003

10. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: A systematic review and meta-analysis. *BMJ* (2018) 362:k3519. doi: 10.1136/bmj.k3519

11. Chun FK, Epstein JI, Ficarra V, Freedland SJ, Montironi R, Montorsi F, et al. Optimizing performance and interpretation of prostate biopsy: A critical analysis of the literature. *Eur Urol.* (2010) 58:851–64. doi: 10.1016/j.eururo.2010.08.041

12. Pye H, Singh S, Norris JM, Carmona EL, Stavrinides V, Grey A, et al. Evaluation of PSA and PSA density in a multiparametric magnetic resonance imaging-directed diagnostic pathway for suspected prostate cancer: The INNOVATE trial. *Cancers (Basel).* (2021) 13(8):1985. doi: 10.3390/cancers13081985

13. Chiu ST, Cheng YT, Pu YS, Lu YC, Hong JH, Chung SD, et al. Prostate health index density outperforms prostate health index in clinically significant prostate cancer detection. *Front Oncol* (2021) 11:772182. doi: 10.3389/fonc.2021.772182

14. Lundgren PO, Kjellman A, Norming U, Gustafsson O. Association between one-time prostate-specific antigen (PSA) test with free/total PSA ratio and prostate cancer mortality: A 30-year prospective cohort study. *Bju Int* (2021) 128:490–6. doi: 10.1111/bju.15417

15. Hirahara T, Arigami T, Yanagita S, Matsushita D, Uchikado Y, Kita Y, et al. Combined neutrophil-lymphocyte ratio and platelet-lymphocyte ratio predicts chemotherapy response and prognosis in patients with advanced gastric cancer. *BMC Cancer.* (2019) 19:672. doi: 10.1186/s12885-019-5903-y

16. Robinson AV, Keeble C, Lo M, Thornton O, Peach H, Moncrieff M, et al. The neutrophil-lymphocyte ratio and locoregional melanoma: A multicentre cohort study. *Cancer Immunol Immunother.* (2020) 69:559–68. doi: 10.1007/s00262-019-02478-7

17. Li S, Xu W, Liao M, Zhou Y, Weng J, Ren L, et al. The significance of gamma-glutamyl transpeptidase to lymphocyte count ratio in the early postoperative recurrence monitoring and prognosis prediction of AFP-negative hepatocellular carcinoma. *J Hepatocell Carcinoma.* (2021) 8:23–33. doi: 10.2147/JHC.S286213

18. Cakir N, Koc AN. Gamma-glutamyl transpeptidase-platelet ratio, systemic immune inflammation index, and system inflammation response index in invasive aspergillosis. *Rev Assoc Med Bras (1992).* (2021) 67:1021–5. doi: 10.1590/1806-9282.20210475

19. Zhang X, Wang D, Chen Z, Guo N, Wang W, Xiong C, et al. Red cell distribution width-to-lymphocyte ratio: A novel predictor for HBV-related liver cirrhosis. *Med (Baltimore).* (2020) 99:e20638. doi: 10.1097/MD.0000000000020600

20. Lehmann F, Schenk LM, Bernstock JD, Bode C, Borger V, Gessler FA, et al. Elevated red cell distribution width to platelet ratio is associated with poor

prognosis in patients with spontaneous, deep-seated intracerebral hemorrhage. *Front Neurol* (2021) 12:751510. doi: 10.3389/fneur.2021.751510

21. Song S, Zhang Y, Qiao X, Duo Y, Xu J, Peng Z, et al. ALT/AST as an independent risk factor of gestational diabetes mellitus compared with TG/HDL-c. *Int J Gen Med* (2022) 15:115–21. doi: 10.2147/IJGM.S332946

22. Reedy DW, Loo AT, Levine RA. AST/ALT ratio > or = 1 is not diagnostic of cirrhosis in patients with chronic hepatitis c. *Dig Dis Sci* (1998) 43:2156–9. doi: 10.1023/a:1018888021118

23. Kather JN, Pearson AT, Halama N, Jager D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* (2019) 25:1054–6. doi: 10.1038/s41591-019-0462-y

24. Diller GP, Kempny A, Babu-Narayan SV, Henrichs M, Brida M, Uebing A, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients. *Eur Heart J* (2019) 40:1069–77. doi: 10.1093/eurheartj/ehy915

25. Wei L, Huang Y, Chen Z, Lei H, Qin X, Cui L, et al. Artificial intelligence combined with big data to predict lymph node involvement in prostate cancer: A population-based study. *Front Oncol* (2021) 11:763381. doi: 10.3389/fonc.2021.763381

26. Nan LB, Yin XT, Gao JP. Significant diagnostic value of (FPSA)/Prostate-specific antigen density (PSAD) and (F/T)/PSAD for prostate cancer of the chinese population in a single institution. *Med Sci Monit* (2019) 25:8345–51. doi: 10.12659/MSM.916900

27. Zheng Y, Huang Y, Cheng G, Zhang C, Wu J, Qin C, et al. Developing a new score system for patients with PSA ranging from 4 to 20 ng/mL to improve the accuracy of PCa detection. *Springerplus.* (2016) 5:1484. doi: 10.1186/s40064-016-3176-3

28. Wu Q, Li F, Yin X, Gao J, Zhang X. Development and validation of a nomogram for predicting prostate cancer in patients with PSA </= 20 ng/mL at initial biopsy. *Med (Baltimore).* (2021) 100:e28196. doi: 10.1097/MD.0000000000028196

29. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med* (2018) 15:e1002699. doi: 10.1371/journal.pmed.1002699

30. Deo RC. Machine learning in medicine. *Circulation.* (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593

31. Komura D, Ishikawa S. Machine learning approaches for pathologic diagnosis. *Virchows Arch* (2019) 475:131–8. doi: 10.1007/s00428-019-02594-w

32. Hajiloo M, Rabiee HR, Anooshahpour M. Fuzzy support vector machine: An efficient rule-based classification technique for microarrays. *BMC Bioinf* (2013) 14 Suppl 13:S4. doi: 10.1186/1471-2105-14-S13-S4

33. Li W, Wang J, Liu W, Xu C, Li W, Zhang K, et al. Machine learning applications for the prediction of bone cement leakage in percutaneous vertebroplasty. *Front Public Health* (2021) 9:812023. doi: 10.3389/fpubh.2021.812023

34. Wang H, Fan T, Yang B, Lin Q, Li W, Yang M. Development and internal validation of supervised machine learning algorithms for predicting the risk of surgical site infection following minimally invasive transforaminal lumbar interbody fusion. *Front Med (Lausanne).* (2021) 8:771608. doi: 10.3389/fmed.2021.771608

35. Cao Y, Giovannucci E. Obesity and prostate cancer. *Recent Results Cancer Res* (2016) 208:137–53. doi: 10.1007/978-3-319-42542-9_8

36. Allott EH, Masko EM, Freedland SJ. Obesity and prostate cancer: Weighing the evidence. *Eur Urol.* (2013) 63:800–9. doi: 10.1016/j.eururo.2012.11.013

37. Vidal AC, Oyekunle T, Howard LE, De Hoedt AM, Kane CJ, Terris MK, et al. Obesity, race, and long-term prostate cancer outcomes. *Cancer-Am Cancer Soc* (2020) 126:3733–41. doi: 10.1002/cncr.32906

38. Graff RE, Ahearn TU, Pettersson A, Ebot EM, Gerke T, Penney KL, et al. Height, obesity, and the risk of TMPRSS2:ERG-defined prostate cancer. *Cancer Epidemiol Biomarkers Prev* (2018) 27:193–200. doi: 10.1158/1055-9965.EPI-17-0547

39. Zhou L, Wang SB, Chen SG, Qu Q, Rui JA. Prognostic value of ALT, AST, and AAR in hepatocellular carcinoma with b-type hepatitis-associated cirrhosis after radical hepatectomy. *Clin Lab* (2018) 64:1739–47. doi: 10.7754/Clin.Lab.2018.180532

40. Chen SL, Li JP, Li LF, Zeng T, He X. Elevated preoperative serum alanine Aminotransferase/Aspartate aminotransferase (ALT/AST) ratio is associated with better prognosis in patients undergoing curative treatment for gastric adenocarcinoma. *Int J Mol Sci* (2016) 17(6):911. doi: 10.3390/ijms17060911

41. Johansen MJ, Gade J, Stender S, Frithioff-Bojsoe C, Lund M, Chabanova E, et al. The effect of overweight and obesity on liver biochemical markers in children and adolescents. *J Clin Endocrinol Metab* (2020) 105(2):dgz010. doi: 10.1210/clinem/dgz010

42. Maldonado-Hernandez J, Saldana-Davila GE, Pina-Aguero MI, Nunez-Garcia BA, Lopez-Alarcon MG. Association between plasmatic ceramides profile and AST/ALT ratio: C14:0 ceramide as predictor of hepatic steatosis in adolescents independently of obesity. *Can J Gastroenterol Hepatol* (2017) 2017:3689375. doi: 10.1155/2017/3689375

43. Yao D, Yang S, Wang Y, Bian K, Yang W, Wang D, et al. An ALP-activatable and mitochondria-targeted probe for prostate cancer-specific bimodal imaging and aggregation-enhanced photothermal therapy. *Nanoscale.* (2019) 11:6307–14. doi: 10.1039/c9nr00913b

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 26 August 2022
DOI 10.3389/fonc.2022.948662

OPEN ACCESS

EDITED BY
Jian Lu,
Peking University Third Hospital, China

REVIEWED BY
Chunxia Su,
Shanghai Pulmonary Hospital, China
Kaustav Bera,
Maimonides Medical Center,
United States

*CORRESPONDENCE
Dong Xu
xudong@zjcc.org.cn
Jincao Yao
yaojc@zjcc.org.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 20 May 2022
ACCEPTED 08 August 2022
PUBLISHED 26 August 2022

CITATION
Wang K, Chen P, Feng B, Tu J, Hu Z, Zhang M, Yang J, Zhan Y, Yao J and Xu D (2022) Machine learning prediction of prostate cancer from transrectal ultrasound video clips. *Front. Oncol.* 12:948662. doi: 10.3389/fonc.2022.948662

Machine learning prediction of prostate cancer from transrectal ultrasound video clips

Kai Wang[1†], Peizhe Chen[2†], Bojian Feng[3,4†], Jing Tu[1], Zhengbiao Hu[1], Maoliang Zhang[1], Jie Yang[1], Ying Zhan[1], Jincao Yao[3,4,5*] and Dong Xu[3,4,5,6*]

[1]Department of Ultrasound, The Affiliated Dongyang Hospital of Wenzhou Medical University, Dongyang, China, [2]College of Optical Science and Engineering, Zhejiang University, Hangzhou, China, [3]Department of Ultrasound, Cancer Hospital of the University of Chinese Academy of Sciences, Zhejiang Cancer Hospital, Hangzhou, China, [4]Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, China, [5]Key Laboratory of Head & Neck Cancer Translational Research of Zhejiang Province, Hangzhou, China, [6]Zhejiang Provincial Research Center for Cancer Intelligent Diagnosis and Molecular Technology, Hangzhou, China

Objective: To build a machine learning (ML) prediction model for prostate cancer (PCa) from transrectal ultrasound video clips of the whole prostate gland, diagnostic performance was compared with magnetic resonance imaging (MRI).

Methods: We systematically collated data from 501 patients—276 with prostate cancer and 225 with benign lesions. From a final selection of 231 patients (118 with prostate cancer and 113 with benign lesions), we randomly chose 170 for the purpose of training and validating a machine learning model, while using the remaining 61 to test a derived model. We extracted 851 features from ultrasound video clips. After dimensionality reduction with the least absolute shrinkage and selection operator (LASSO) regression, 14 features were finally selected and the support vector machine (SVM) and random forest (RF) algorithms were used to establish radiomics models based on those features. In addition, we creatively proposed a machine learning models aided diagnosis algorithm (MLAD) composed of SVM, RF, and radiologists' diagnosis based on MRI to evaluate the performance of ML models in computer-aided diagnosis (CAD). We evaluated the area under the curve (AUC) as well as the sensitivity, specificity, and precision of the ML models and radiologists' diagnosis based on MRI by employing receiver operator characteristic curve (ROC) analysis.

Results: The AUC, sensitivity, specificity, and precision of the SVM in the diagnosis of PCa in the validation set and the test set were 0.78, 63%, 80%; 0.75, 65%, and 67%, respectively. Additionally, the SVM model was found to be superior to senior radiologists' (SR, more than 10 years of experience) diagnosis based on MRI (AUC, 0.78 vs. 0.75 in the validation set and 0.75 vs. 0.72 in the test set), and the difference was statistically significant ($p < 0.05$).

Conclusion: The prediction model constructed by the ML algorithm has good diagnostic efficiency for prostate cancer. The SVM model's diagnostic

efficiency is superior to that of MRI, as it has a more focused application value. Overall, these prediction models can aid radiologists in making better diagnoses.

KEYWORDS

artificial intelligence, prostate cancer, ultrasound, machine learning, support vector machine

Introduction

Prostate cancer (PCa) is one of the most common cancers in males, and its prevalence has increased at an alarming rate over the last several decades (1). According to GLOBOCAN 2020, in 2020 there were approximately 1,414,259 new cases of PCa and 375,304 PCa-related deaths worldwide, with a particularly high prevalence in developed countries (2). The early clinical manifestations of prostate cancer are sufficiently nonspecific that patients often ignore it in its early phases and therefore only seek treatment when it has already developed. Therefore, early diagnosis of PCa is crucial. Prostate-specific antigen (PSA) testing, digital rectal examinations (DRE), and transrectal ultrasonography (TRUS) guided prostate system biopsies are the most used PCa screening methods in clinics (3, 4), but these diagnostic tools may still lead to a certain degree of overdiagnosis (5).

In the past decade, the role of MRI in the diagnosis of prostate cancer stages has significantly developed. The introduction of coil imaging in the rectum and the advent of some basic techniques, such as magnetic resonance spectrum imaging, dynamic contrast-enhanced MRI, and diffusion-weighted imaging (DWI), have improved the diagnostic accuracy of MRI and its potential to improve the treatment decision-making process (6). However, it must be emphasized that multiparametric magnetic resonance imaging (mpMRI) has been evaluated only in patients in whom the risk of clinically significant PCa was judged sufficiently high to warrant biopsy. Therefore, a prebiopsy mpMRI must not be used as an initial screening tool. Indeed, based on its low specificity, mpMRI in very low-risk patients would result in an increase in false-positive findings and subsequent unnecessary biopsies (5). Another classical imaging diagnosis of PCa has largely relied on ultrasonography (US), including transrectal ultrasound (TRUS), contrast-enhanced ultrasonography (7), and ultrasound elastography (8). However, sonologists' evaluations of tumor tissue have primarily relied on semantic features from the visual perspective, which is an approach that misses many image features that represent tumor heterogeneity. Therefore, early accurate diagnosis of prostate cancer remains a clinical challenge.

As a new frontier, ML-based radiomics could extract many quantitative features from encrypted digital images, which could then be used to deeply mine the biological information of tumors and analyze the heterogeneity of tumors, thus aiding clinical decision making (9). However, it has been reported that the ML ultrasound diagnostic model is rarely used to evaluate PCa because prostate cancers located in the central zone are often difficult to detect visually—they are confused with the hypoechoic endoglandular background tissue. Plus, the application of ML methods on prostate cancer prediction is mostly based on static transrectal images, which cannot fully display the patient's tissue information, meaning that the application of ML methods for prostate cancer prediction based on transrectal video clips remains problematic.

Accordingly, we adopted ML models based on whole prostate transrectal ultrasound video clips on PCa prediction. The ML algorithms were better at forming predictions because they could use the ultrasound video clips to make a prediction based on global information from whole prostate transrectal tissue. To test the performance of the ML algorithms, we compared their diagnostic results with radiologists' diagnosis based on MRI, finding that the SVM algorithm adopted in this paper had better performance in terms of PCa prediction. In addition, we creatively proposed a machine learning models aided diagnosis algorithm (MLAD) to evaluate the performance of ML models in computer-aided diagnosis (CAD).

Materials and methods

Local ethics committees approved the study (2022-YX-047).

Abbreviations: ML, machine learning; PCa, prostate cancer; MRI, magnetic resonance imaging; SVM, support vector machine; RF, random forest; MLAD, machine learning models aided diagnosis algorithm; CAD, computer-aided diagnosis; JR, junior radiologist; SR, senior radiologist; LASSO, the least absolute shrinkage and selection operator; DWI, diffusion-weighted imaging; AUC, area under the curve; ROC, receiver operator characteristic curve; US, ultrasonography; TRUS, transrectal ultrasound; ICCs, intra- and inter-class correlation coefficients; ROI, region of interest.

Study participants

We obtained the ultrasound video clips data from Dongyang Hospital, which is affiliated with Wenzhou Medical University. From January 2021 to December 2021, we recruited 276 patients with PCa and 225 patients with benign lesions of the prostate, which included benign prostatic hyperplasia, fibromuscular tissue, atypical glandular tissue, and chronic prostatitis.

The inclusion criteria were as follows: (a) was an elderly male (aged above 55); (b) had solid prostate masses found by digital rectal examination, TRUS, or MRI; (c) had undergone prostate biopsy or surgery and obtained the pathological diagnosis results; and (d) had not received treatment for prostate diseases before TRUS. In addition, we excluded patients with rectal malformation or rectal surgery who could not be examined by transrectal ultrasound.

Finally, of the 231 patients—113 having PCa and 118 benign lesions—we randomly selected 170 for the purpose of training and validating an ML model, while using the remaining 61 to test derived models. The processes underlying the inclusion and exclusion of study participants are shown in Figure 1.

Video clips acquisition

We collected all ultrasound data related to the prostate using the Esaote MyLab^{TM} ClassC ultrasound machine (Esaote, Genoa, Italy) with the TRT33 Transrectal Biplane Transducer (frequency range 3–13MHz). Four sonologists, each with over ten years of experience in transrectal ultrasound, performed all ultrasound scans. First, they placed a condom on the TRT33 probe, then inserted it into the rectum, adjusted the probe depth, and rotated the probe for multidirectional prostate examination. Second, they scanned the entire transverse section of the prostate grayscale ultrasound, before scanning the prostate from top to bottom and storing 10 seconds of video clips.

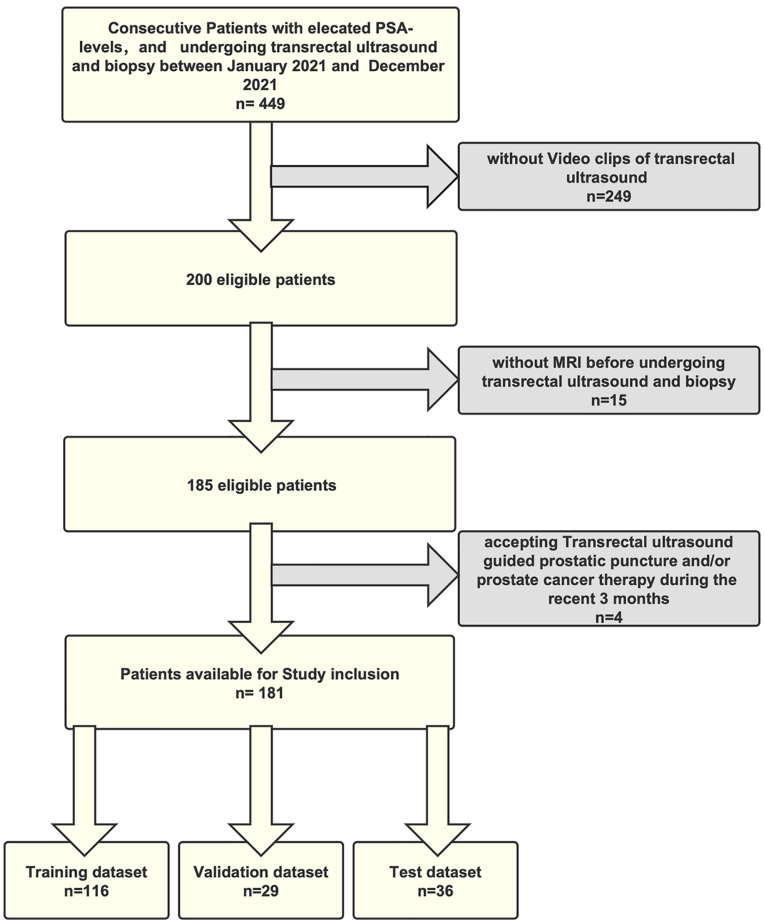


FIGURE 1
The flowchart of inclusion and exclusion of the study population.

Manual segmentation

For manual segmentation, we loaded the video clips into the 3D Slicer v.4.11. Two sonologists (S1 and S2, each with more than five years of TRUS prostate diagnosis experience) manually segmented the region of interest (ROI) from the prostate. They were blinded to the MRI and pathological results. We drew the entire prostate gland as an ROI from every video frame and used intra- and inter-class correlation coefficients (ICCs) to evaluate the reproducibility of radiomics feature extraction. First, S1 and S2 separately segmented video clips of 30 randomly selected patients, and then, two weeks later, S2 segmented images of the 30 patients once more. After that, S1 performed the remaining video clip segmentation. We only included the features with an ICC value equal to or higher than 0.8 that indicated excellent reproducibility in the other feature selection process.

Feature extraction and selection

We performed extractions of radiomics features by using a radiomics extension of a 3D Slicer software, SlicerRadomics (Version 3.0.1) (10). For this, we extracted 851 radiomics features from each patient, including shape features (14), first-order statistical features (18), gray-level co-occurrence matrix features (24), gray-level dependence matrix features (14), gray-level run-length matrix features (16), gray-level size zone matrix features (16), neighborhood gray-tone difference matrix features (5), and wavelet-based features (744).

We performed feature selection using programs written in Python (Version 3.8.8, Python Software Foundation). First, according to pathological results, we divided all data into benign and malignant groups and inserted labels 0 and 1 into the data. Second, we used the two independent samples t-test and Mann-Whitney U test to test all the features, before deleting the features in the benign and malignant groups that failed to meet either of the first two tests. Third, the least absolute shrinkage and selection operator (LASSO) regression selected features in the training and validation set. We excluded the features with zero variance using the variance filtering method. Fourth, we performed the LASSO method for further dimensionality reduction of the features and selected the most valuable features (11). We then repeated the 10-fold cross-validation on training and validation set process 100,000 times to obtain the optimal value of parameter λ, which we introduced into the LASSO method to calculate the regression coefficients of each feature. Finally, we selected the features with non-zero coefficients.

Machine learning

Python scikit-learn 0.24.2 package (12) was used to support vector machine modeling and evaluation. We randomly divided the training and validation set into the training set and the validation set at a ratio of 8:2. First, we used a Gaussian kernel support vector machine (SVM) model and a random forest (RF) model to classify features in the training set and established two nonlinear classifiers. In the SVM classifier, kernel size parameters (γ, gamma) and regularization parameters (C, cost) of the SVM kernel function were optimized. We then selected the parameters with the best performance through 10-fold cross-validation on the training set. In the RF classifier, the number of estimators (n_estimator) was optimized through 10-fold cross-validation on the training set. Finally, we applied the SVM model and RF model to the validation set and test set.

MRI results collection

Participants underwent prostate MRIs with a 1.5T Siemens Magnetom Avanto magnetic resonance scanner (Erlangen, Germany), including the standard T2-weighted MRI, T1-weighted MRI, and diffusion-weighted MRI. We included the MRI results of the validation and test set in the analysis. Two junior radiologists (each with less than 5 years of experience) and two senior radiologists (each with more than 10 years of experience) reviewed each case in order to provide an independent diagnosis.

Machine learning models aided diagnosis algorithm

Using the SVM model and RF model, we can obtain the probabilities of PCa separately. To evaluate the performance of ML models in CAD, we proposed a machine learning models aided diagnosis algorithm (MLAD) (Equation 1) which integrates the prediction performance of SVM, RF, and radiologists' diagnosis. In this algorithm, we chose SVM as the main model and RF as the sub model.

$$\begin{cases} S_{SVM} = |0.5 - P_{SVM}| \\ S_{RF} = |0.5 - P_{RF}| \\ S_{MLAD} = \frac{S_{SVM}}{0.5} * P_{SVM} + \frac{0.5 - S_{SVM}}{0.5} * \left(\frac{S_{RF}}{0.5} * P_{RF} + \frac{0.5 - S_{RF}}{0.5} * V_R \right) \end{cases} \quad (1)$$

P_{SVM}: probability of PCa from SVM model; P_{RF}: probability of PCa from RF model; V_R: result value of PCa from radiologists' diagnosis (0: benign, 1: malignant); S_{SVM}: prediction confidence score of SVM; S_{RF}: prediction confidence score of RF

Statistical analysis

We used SPSS 25.0 software to conduct a statistical analysis and tested the normality of continuous variables using the Levene test. We analyzed continuous variables obeying a normal distribution by

using the independent samples t-test and analyzed those not obeying a normal distribution by using the Mann-Whitney U test. We then compared categorical variables using the chi-square test. Unless otherwise specified, we expressed the continuous variables as median (standard deviation, SD) and the categorical variables as n (%). We also evaluated the area under the curve (AUC) as well as the sensitivity, specificity, and precision of the SVM model and MRI in diagnosing prostate cancer by employing receiver operator characteristic curve (ROC) analysis. $P< 0.05$ indicated a significant difference. The overall flowchart of the study is outlined in Figure 2.

Results

Clinicopathological features of the patients

The clinicopathological features in the training set, validation set, and test set are shown in Table 1. The average ages of patients in the training set, validation set, and test set were 72.02, 71.21, and 69.64, respectively, while the respective mean PSA values were 19.91, 22.77, and 46.98. The number of benign lesions were 83 (48.8%), 15 (44.1%), and 30 (49.2%). The number of PCa were 87 (51.2%), 19 (55.9%), and 31 (50.8%). We found no significant differences between the validation and test set in terms of age, PSA, or pathological results ($p > 0.05$).

Feature selection

From each patient, we extracted 851 features from the ultrasound video clips using the LASSO regression model in the training set. In the LASSO model, we repeated the 10-fold cross-validation process 100,000 times in order to generate the optimal penalization coefficient lambda (λ).

Finally, we chose a λ value of 0.029470517025518096. After dimensionality reduction with LASSO regression, 14 features were selected, consisting of original (3) and wavelet features (11). The subset of features ultimately selected by the LASSO algorithm is shown in Table 2. Figure 3, meanwhile, shows the selection of significant parameters in features in the training set and the definition of the linear predictor, while Figure 4 shows the generation of the optimal penalization coefficient lambda.

Modeling and effectiveness

We first selected the SVM algorithm to establish the radiomics model based on the selected 14 features. Traditionally, the prediction performance has been optimized for the following parameters: C, gamma, and the shape of the kernel. We then constructed a pipeline with two steps: a scaling step and an SVM step. It is best to scale data before passing them to an SVM. Next, we varied the relevant RBF parameters, C and gamma, logarithmically, varying by one order of magnitude at a time. We used a 10-fold cross-validation scheme. Finally, we identified the best SVM estimator (C=45.20, gamma=0.001) and stored it as an SVM model. As with the SVM model, in the RF algorithm, we varied the number of estimators by one order of magnitude at a time in order to obtain the best RF model (n_estimators=10000). The AUC, sensitivity, specificity, and precision of the SVM and RF model in the diagnosis of PCa in the validation set and the test set were as follows: (1) SVM results: 0.78, 63% (95%CI: 0.38–0.83), 80% (95%CI: 0.51–0.95),

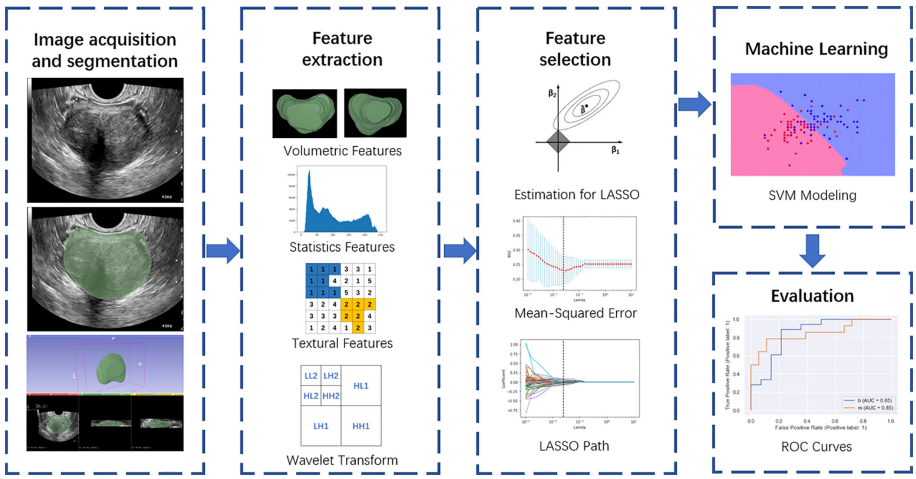


FIGURE 2
Overall flow chart of the study, including image acquisition and segmentation, feature extraction, feature selection, machine learning, and evaluation.

TABLE 1 Characteristics of patients in the training, validation and test datasets.

	Training set	Validation set	Test set	*P* value
Age(y)*	72.02±8.721	71.21±6.246	69.64±8.262	0.161
PSA(ng/mL)*	19.91±44.56	22.77±65.51	46.98±114.89	0.065
Pathology				0.871
No.of Benign(-)(%)	83(48.8%)	15(44.1%)	30(49.2%)	
BPH	68(40%)	13(38.3%)	25(41%)	
BPH & prostatitis	11(6.4%)	1(2.9%)	3(4.9%)	
BPH & BCH	1(0.6%)	0	2(3.2)	
BPH & LGIN	3(1.8%)	1(2.9%)	0	
No. of Pca(+)(%)	87(51.2%)	19(55.9%)	31(50.8%)	
GS6	30(17,6%)	9(26.5%)	14(23%)	
GS7	38(22.4%)	7(20.6%)	8(13.1%)	
GS8	10(5.9%)	2(5.9%)	6(9.8%)	
GS>=8	9(5.3%)	1(2.9%)	3(4.9%)	

BPH, benign prostatic hyperplasia; BCH, basal cell hyperplasia; LGIN. low-grade intraepithelial neoplasia.
*Data are expressed as mean ± standard deviation.
p< 0.05 indicates significant differences in patients' clinicopathological features in the validation and test sets.

TABLE 2 The subset of radiomics features ultimately selected by the LASSO algorithm.

Feature	Image type	Feature Class	Feature Name	LASSO coefficients
1	Original	Firstorder	Range	0.049561
2	Original	glcm	ClusterProminence	0.004193
3	Original	glszm	ZoneEntropy	0.011709
4	Wavelet-LHL	firstorder	Skewness	-0.055034
5	Wavelet-LHL	glcm	ClusterShade	-0.025479
6	Wavelet-LHL	glcm	Correlation	0.010685
7	Wavelet-LHH	gldm	LargeDependenceLowGrayLevelEmphasis	0.018646
8	Wavelet-HLL	glszm	GrayLevelNonUniformity	-0.073279
9	Wavelet-HHH	firstorder	Median	-0.050800
10	Wavelet-HHH	glcm	ClusterShade	0.025124
11	Wavelet-HHH	gldm	LargeDependenceLowGrayLevelEmphasis	0.039665
12	Wavelet-LLL	glszm	LargeAreaHighGrayLevelEmphasis	-0.026230
13	Wavelet-LLL	glszm	SizeZoneNonUniformityNormalized	0.021580
14	Wavelet-LLL	glszm	SmallAreaHighGrayLevelEmphasis	0.056598

First-order features describe the distribution of voxel intensities within the image region defined by the mask through commonly used and basic metrics. GLCM features describe the second-order joint probability function of an image region constrained by the mask. They are defined as P. GLDM features quantify gray-level dependencies in an image, and GLSZM features quantify gray-level zones in an image.

80% (0.51–0.95); 0.75, 65% (95%CI: 0.45–0.80), 67% (95%CI: 0.47–0.82), 67% (95%CI: 0.47–0.82); (2) RF results: 0.77, 63% (95%CI: 0.39–0.83), 87% (95%CI: 0.58–0.98), 86% (95%CI: 0.56–0.97); 0.69, 45% (95%CI: 0.28–0.64), 93% (95%CI: 0.76–0.99), 88% (95%CI: 0.60–0.98). For comparison, two junior radiologists (each with less than 5 years of experience) and two senior radiologists (each with more than 10 years of experience) respectively gave their independent diagnosis of PCa based on MRI results. The AUC, sensitivity, specificity, and precision of two radiologists in the diagnosis of PCa based on MRI were as follows: (1) JR results: 0.65, 63% (95%CI: 0.39–0.83), 67% (95%CI: 0.39–0.87), 71% (95%CI: 0.44–0.87); 0.65, 71% (95%CI: 0.52–0.85), 60% (95%CI: 0.41–0.77), 65% (95%CI: 0.46–0.80); (2) SR results: 0.75, 63% (95%CI: 0.39–0.83), 87% (95%CI: 0.58–0.98), 86% (95%CI: 0.56–0.97); 0.72, 61% (95%CI: 0.42–0.78), 83% (95%CI: 0.65–0.93), 0.79 (95%CI: 0.57–0.92) (Table 3). According to the statistical results, the SVM model was superior to radiologists' diagnosis based on MRI (AUC, 0.78 vs. 0.65/0.75 and 0.75 vs. 0.65/0.72) (Figure 5), and the results of the SVM model and SR were statistically significant ($p< 0.05$). To evaluate the performance of ML models in CAD, we integrated the SVM model and RF model with JR and SR diagnosis through MLAD separately (SVM+RF+JR and SVM+RF+SR). The AUC, sensitivity, specificity, and precision of the MLAD model in the diagnosis of PCa in the validation set and the test set were as follows: (1) SVM+RF+JR: 0.8, 74% (95%CI: 0.49–0.90), 87%

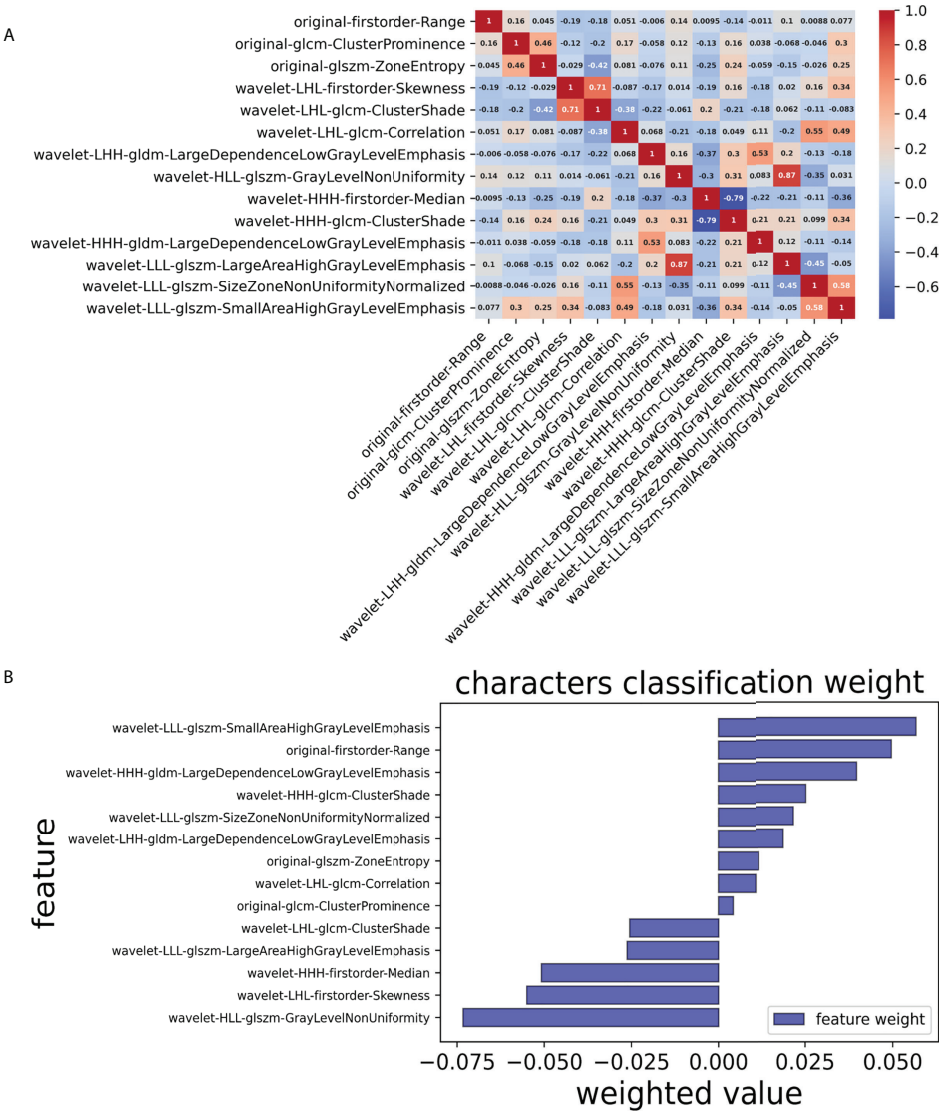


FIGURE 3
Selection of significant parameters in features in the training set and definition of the linear predictor. **(A)** Spearman's correlation coefficients were calculated for the fourteen selected features. **(B)** Characters classification weight of the features.

(95%CI: 0.58–0.98), 88% (95%CI: 0.60–0.98); 0.72, 68% (95%CI: 0.49, 0.83), 70% (95%CI: 0.50, 0.85), 70% (95%CI: 0.50, 0.85); (2) SVM+RF+SR: 0.86, 74% (95%CI: 0.49–0.90), 93% (95%CI: 0.66–0.99), 93% (95%CI: 0.66–0.99); 0.81, 81% (95%CI: 0.62, 0.92), 80% (95%CI: 0.61–0.92), 81% (95%CI: 0.62–0.92). According to the statistical results, the MLAD model with senior radiologists' diagnosis (SVM+RF+SR) was superior to senior radiologists' diagnosis based on MRI and the SVM model (AUC, 0.85 vs. 0.75/0.78 and 0.81 vs. 0.72/0.75) (Figure 5), and the results were statistically significant ($p< 0.05$). The results thus demonstrated that the SVM model and RF model can improve the predictive performance of PCa through MLAD.

Discussion

ML-based radiomics transforms visual image information into in-depth feature quantitative data, extracts a large amount of image characteristic information from medical images, and constructs pre-measurement models based on feature information (13, 14). In this study, we carried out feature extraction of the prostate creatively from ultrasound video clips. The prediction model constructed by the ML algorithm has good diagnostic efficiency in PCa, and, compared with the SVM model with an MRI, the diagnostic efficiency is better and has a more specific application value.

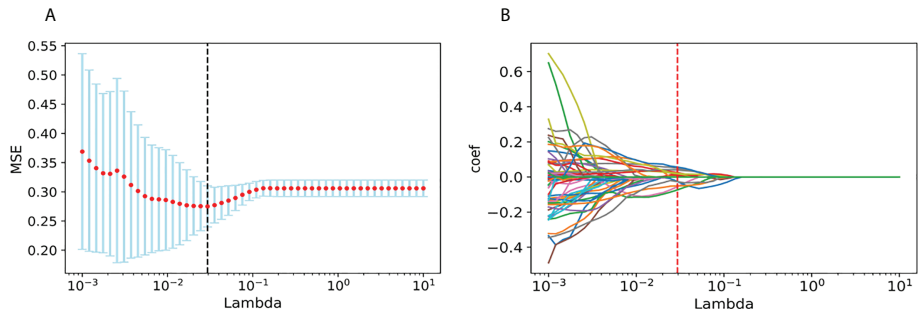


FIGURE 4
Generation of the optimal penalization coefficient lambda. **(A)** Ten-time cross-validation for tuning parameter selection in the LASSO model. **(B)** LASSO coefficient solution path for the 14 features.

TABLE 3 Diagnostic performance of machine learning model and MRI on a per-lesion basis.

Dataset and Method	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	AUC	*P* Value	Kappa
SVM Model						
Validation	0.63 (0.38-0.83)	0.80 (0.51-0.95)	0.80 (0.51-0.95)	0.78	0.012	0.42
Test	0.65 (0.45-0.80)	0.67 (0.47-0.82)	0.67 (0.47-0.82)	0.75	0.015	0.312
RF Model						
Validation	0.63 (0.37-0.83)	0.87 (0.58-0.98)	0.86 (0.56-0.97)	0.77	0.003	0.481
Test	0.45 (0.28-0.64)	0.93 (0.76-0.99)	0.88 (0.60-0.98)	0.69	0.001	0.382
MRI-JR						
Validation	0.63 (0.39-0.83)	0.67 (0.39-0.87)	0.71 (0.44-0.87)	0.65	0.084	0.294
Test	0.71 (0.52-0.85)	0.60 (0.41-0.77)	0.65 (0.46-0.80)	0.65	0.015	0.31
MRI-SR						
Validation	0.63 (0.39-0.83)	0.87 (0.58-0.98)	0.86 (0.56-0.97)	0.75	0.003	0.481
Test	0.61 (0.42-0.78)	0.83 (0.65-0.93)	0.79 (0.57-0.92)	0.72	0.0003	0.445
SVM+RF+JR						
Validation	0.74 (0.49-0.90)	0.87 (0.58-0.98)	0.88 (0.60-0.98)	0.8	0.000464	0.591
Test	0.68 (0.49-0.83)	0.70 (0.50-0.85)	0.70 (0.50-0.85)	0.72	0.003	0.377
SVM+RF+SR						
Validation	0.74 (0.49-0.90)	0.93 (0.66-0.99)	0.93 (0.66-0.99)	0.85	0.00009	0.652
Test	0.81 (0.62-0.92)	0.80 (0.61-0.92)	0.81 (0.62-0.92)	0.81	0.000002	0.606

SVM model, support vector machine model; RF model, random forest model; MRI-JR, junior radiologists' (less than 5 years of experience) diagnosis based on MRI; MRI-SR, senior radiologists' (more than 5 years of experience) diagnosis based on MRI. $p < 0.05$ indicates a significant difference in the discrimination of the SVM model and MRI diagnosis.

TRUS is widely used in clinical practice because it is safe, radiation-free, inexpensive, and easy to perform (15, 16). The outline of the prostate is usually clearly displayed, and the boundary between the isoechoic peripheral and hypoechoic central zone is demarcated. Approximately 70% of PCa is located in the peripheral area, and most PCa is hypoechoic. However, tumors in the central location are often difficult to detect because they are confused with the low-echoic endoglandular background tissue. Thus, TRUS has only moderate accuracy in PCa detection in the general population (17). Therefore, we used the video clips of the prostate to serve as the feature extraction data in order to avoid losing a key portion of the information.

By extracting high-throughput data and establishing an efficient and stable prediction model, radiomics can provide an auxiliary diagnosis for clinical practice. Features selection is the key to ML research. However, data redundancy and over-fitting will occur if the high-throughput feature extraction is not selected (18–20). Jin et al. used the ML method to predict lymph node metastasis of early cervical cancer and extracted 106 imaging omics features from lymph node ultrasound images. Through a combination of LASSO and ridge regression, they selected the key features from the high-dimensional features to avoid overfitting. They then selected six features for classification research, which represented the

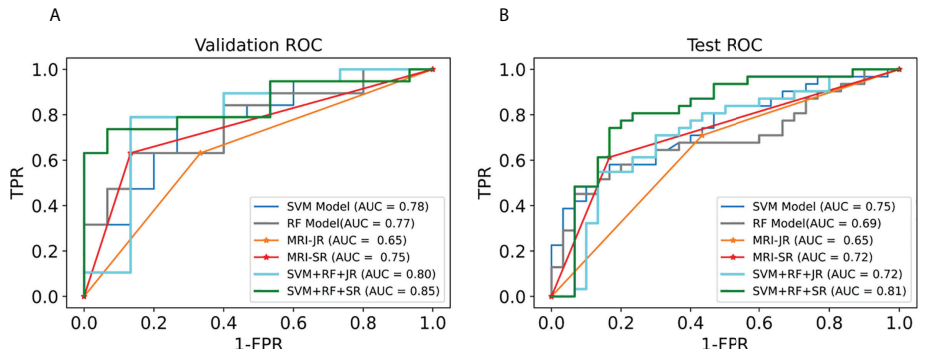


FIGURE 5
Comparison of ROC between the ML models and MRI in the validation set and test set. **(A)** shows the ROC curves of the validation set. **(B)** shows the ROC curves of the test set. (MRI-JR: junior radiologists' diagnosis based on MRI, MRI-SR: senior radiologists' diagnosis based on MRI).

texture complexity of tumors and correlated with the high degree of tumor heterogeneity (21). By contrast, in our study, we extracted a total of 851 features. We used the LASSO algorithm to filter all the features and retained only the 14 non-zero features with a solid correlation with PCa. The results show that the AUC of the prediction model in the PCa of the training set, validation set, and test set were respectively 0.82, 0.78, and 0.75. The difference was not statistically significant (p >0.05) and it has been shown that the feature selection method used in this study can effectively restrain data overfitting.

Out of 14 selected image omics features (22), one was taken from the first-order range, i.e., the range of gray values in the ROI; one was from the gray-level co-occurrence matrix (GLCM), which describes the second-order joint probability function of an image region constrained by the mask; one was from the gray-level size zone matrix (GLSZM), which quantifies gray level zones in an image; and the remaining 11 were taken from wavelet (23), subsets of texture features. Among the selected radiomics features, texture features based on wavelet account for the majority, which indicates that texture features have a good classification function. Furthermore, they are related to the composition of heterogeneous cells with noticeable molecular and microenvironmental differences in malignant tumors, indicating that the texture characteristics of tumors are highly correlated with heterogeneity (21).

The commonly used modeling methods of radiomics are mainly divided into statistical and ML-based methods. ML approaches are then typically subdivided into supervised and unsupervised learning, with SVM and RF being the most widely used approaches in supervised learning (24–28). Previous literature has reported that both SVM and RF show good stability. Specifically, SVM and RF show good diagnostic efficiency in constructing a small sample prediction model (29). The basic principle of SVM is to divide a hyperplane into a given training queue space to distinguish different types of samples (30). In this study, we used SVM and RF to construct a radiomics prediction model. The AUC of SVM in the validation and test set was 0.78 and 0.75, respectively, showing that the model had high diagnostic efficiency and stability.

To evaluate the diagnostic efficacy of the ML prediction SVM model in PCa, we compared the SVM model with MRI diagnosis. In Rosenkrantz et al.'s study, it was reported that the sensitivity, specificity, and precision of PCa detection using a fusion of T2-weighted images and diffusion-weighted images were 60.8%, 80.3%, and 71.0%, respectively (31). The results of Katahira et al.'s study, meanwhile, which used T2WI and DWI to detect prostate cancer, showed that the sensitivity, specificity, and AUC were 61.2%, 82.6%, and 0.755, respectively (32). In our study, the AUC, sensitivity, specificity, and precision of two radiologists in the diagnosis of PCa based on MRI were 0.65/0.75, 63%/63%, 67%/87%, 71%/86%; 0.65/0.72, 71%/61%, 60%/83%, 65%/79% (the former is the diagnosis results of JR, and the latter is the diagnosis results of SR). The diagnostic power of MRIs in this study was similar to that observed in previous studies. The results show that the SVM model had higher diagnostic efficiency than a diagnosis based on MRI (AUC, 0.78 vs. 0.65/0.75 and 0.75 vs. 0.65/0.72). In the extended experiment, the MLAD model with SR diagnosis (SVM+RF +SR) showed the best performance in statistics, which means the SVM model and RF model can improve radiologists' diagnosis performance.

In spite of its findings, it must be acknowledged that this study suffers from some limitations. The most obvious limitation is that it was a retrospective study with a small sample size, limiting its power and precluding firm statistical conclusions. For example, the CI of AUC sensitivity, specificity, and precision has a large range. The second limitation is that it was a single-center study, and thus we cannot exclude single-centered effects. Finally, we carried out ROI segmentation manually, which is inefficient and may lead to bias among different delineators, thus resulting in the reduced diagnostic capability of the prediction model. Although the radiomics model performed well in this

study, future studies must combine clinical factors closely related to PCa to build a more robust model.

It is rare to use the ML ultrasound diagnostic model to evaluate PCa. Therefore, we aimed to build an ultrasound diagnostic prediction model based on ML to provide a solid theoretical basis for the accurate and individualized treatment of PCa.

Conclusions

In our study, we innovatively used ultrasound video clips instead of images to form a dataset on which we could build an ML model. The ML-based prediction models have good diagnostic efficiency in PCa. In the SVM model, the precision, sensitivity, and specificity are better than that seen in diagnosis based on MRI. Thus, based on our MLAD, the SVM model and RF model can contribute to improve radiologists' diagnosis performance based on MRI; indeed, the MLAD model in conjunction with senior radiologists' diagnosis shows the best performance among all models. In our future work, we intend to combine the ultrasound transverse and longitudinal video clips of the prostate to build a better ML model and use deep learning and neural networks in the ultrasonic diagnosis of prostate cancer. The model developed in our study could contribute to reducing barriers and providing a convenient way for community hospitals to improve PCa diagnosis.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Dongyang People's Hospital. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

KW and DX designed the study. KW, JT, ZH, MZ, and ZW contributed to data acquisition. PC, JY and YZ carried out statistical analysis. PC, BF, JY, JT, and ZH interpreted the results. KW and JY wrote the draft. PC and BF reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Funding

The study was supported in part by the National Natural Science Foundation of China (82071946), the Natural Science Foundation of Zhejiang Province (LZY21F030001 and LSD19H180001), the Medical and Health Research Project of Zhejiang Province (2021KY099 and 2022KY110) and the funds from the University Cancer Foundation *via* the Sister Institution Network Fund at the University of Texas MD Anderson Cancer Center, and the Key Science and Technology Project of Jinhua, Zhejiang Province (2022-3-017).

Conflict of interest

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.

Publisher's note

References

1. Wang TH, Lee CY, Lee TY, Huang HD, Hsu JB, Chang TH. Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers (Basel)* (2021) 13(11):2528. doi: 10.3390/cancers13112528

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca-a Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660

3. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. *Bmj* (2018) 362:k3519. doi: 10.1136/bmj.k3519

4. Vale CL, Fisher D, Kneebone A, Parker C, Pearse M, Richaud P, et al. Adjuvant or early salvage radiotherapy for the treatment of localised and locally advanced prostate cancer: A prospectively planned systematic review and meta-analysis of aggregate data. *Lancet* (2020) 396(10260):1422–31. doi: 10.1016/S0140-6736(20)31952-8

5. Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. part 1: Screening, diagnosis, and local treatment with curative intent. *Eur Urol* (2021) 79(2):243–62. doi: 10.1016/j.eururo.2020.09.042

6. Kim CK, Park BK. Update of prostate magnetic resonance imaging at 3 T. *J Comput Assist Tomogr* (2008) 32(2):163–72. doi: 10.1097/RCT.0b013e3180683b99

7. Tang J, Yang JC, Luo Y, Li J, Li Y, Shi H. Enhancement characteristics of benign and malignant focal peripheral nodules in the peripheral zone of the prostate gland studied using contrast-enhanced transrectal ultrasound. *Clin Radiol* (2008) 63(10):1086–91. doi: 10.1016/j.crad.2007.11.026

8. Correas JM, Tissier AM, Khairoune A, Vassiliu V, Méjean A, Hélénon O, et al. Prostate cancer: Diagnostic performance of real-time shear-wave elastography. *Radiology* (2015) 275(1):280–9. doi: 10.1148/radiol.14140567

9. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin* (2019) 69(2):127–57. doi: 10.3322/caac.21552

10. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* (2017) 77(21):e104–7. doi: 10.1158/0008-5472.CAN-17-0339

11. Ren S, Qi Q, Liu S, Duan S, Mao B, Chang Z, et al. Preoperative prediction of pathological grading of hepatocellular carcinoma using machine learning-based ultrasomics: A multicenter study. *Eur J Radiol* (2021) 143:109891. doi: 10.1016/j.ejrad.2021.109891

12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30. doi: 10.5555/1953048.2078195

13. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to radiomics. *J Nucl Med* (2020) 61(4):488–95. doi: 10.2967/jnumed.118.222893

14. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. *Strahlenther Onkol* (2020) 196(10):879–87. doi: 10.1007/s00066-020-01625-9

15. Kaneko M, Sugano D, Lebastchi AH, Duddalwar V, Nabhani J, Haiman C, et al. Techniques and outcomes of MRI-TRUS fusion prostate biopsy. *Curr Urol Rep* (2021) 22(4):27. doi: 10.1007/s11934-021-01037-x

16. Merriel SWD, Pocock L, Gilbert E, Creavin S, Walter FM, Spencer A, et al. Systematic review and meta-analysis of the diagnostic accuracy of prostate-specific antigen (PSA) for the detection of prostate cancer in symptomatic patients. *BMC Med* (2022) 20(1):54. doi: 10.1186/s12916-021-02230-y

17. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp Oncol (Pozn)* (2015) 19(1a):A68–77. doi: 10.5114/wo.2014.47136

18. Abboud B, Smayra T, Jabbour H, Ghorra C, Abadjian G. Correlations of neck ultrasound and pathology in cervical lymph node of papillary thyroid carcinoma. *Acta Chir Belg* (2020) 120(4):238–44. doi: 10.1080/00015458.2019.1592988

19. Zhang H, Li X, Zhang Y, Huang C, Wang Y, Yang P, et al. Diagnostic nomogram based on intralesional and perilesional radiomics features and clinical factors of clinically significant prostate cancer. *J Magn Reson Imaging* (2021) 53 (5):1550–8. doi: 10.1002/jmri.27486

20. Xian MF, Zheng X, Xu JB, Li X, Chen LD, Wang W. Prediction of lymph node metastasis in rectal cancer: comparison between shear-wave elastography based ultrasomics and MRI. *Diagn Interv Radiol* (2021) 27(3):424–31. doi: 10.5152/dir.2021.20031

21. Jin X, Ai Y, Zhang J, Zhu H, Jin J, Teng Y, et al. Noninvasive prediction of lymph node status for patients with early-stage cervical cancer based on radiomics features from ultrasound images. *Eur Radiol* (2020) 30(7):4117–24. doi: 10.1007/s00330-020-06692-1

22. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* (2018) 19 (6):1236–46. doi: 10.1093/bib/bbx044

23. de Cheveigné A, Nelken I. Filters: When, why, and how (Not) to use them. *Neuron* (2019) 102(2):280–93. doi: 10.1016/j.neuron.2019.02.039

24. Wang W, Zhang JC, Tian WS, Chen LD, Zheng Q, Hu HT, et al. Shear wave elastography-based ultrasomics: Differentiating malignant from benign focal liver lesions. *Abdom Radiol (NY)* (2021) 46(1):237–48. doi: 10.1007/s00261-020-02614-3

25. Wang W, Wu SS, Zhang JC, Xian MF, Huang H, Li W, et al. Preoperative pathological grading of hepatocellular carcinoma using ultrasomics of contrast-enhanced ultrasound. *Acad Radiol* (2021) 28(8):1094–101. doi: 10.1016/j.acra.2020.05.033

26. Twilt JJ, van Leeuwen KG, Huisman HJ, Fütterer JJ, de Rooij M. Artificial intelligence based algorithms for prostate cancer classification and detection on magnetic resonance imaging: A narrative review. *Diagnostics (Basel)* (2021) 11(6) 959. doi: 10.3390/diagnostics11060959

27. Tu SJ, Tran VT, Teo JM, Chong WC, Tseng JR. Utility of radiomic zones for risk classification and clinical outcome predictions using supervised machine learning during simultaneous (11) c-choline PET/MRI acquisition in prostate cancer patients. *Med Phys* (2021) 48(9):5192–201. doi: 10.1002/mp.15064

28. Sushentsev N, Rundo L, Blyuss O, Nazarenko T, Suvorov A, Gnanapragasam VJ, et al. Comparative performance of MRI-derived PRECISE scores and delta-radiomics models for the prediction of prostate cancer progression in patients on active surveillance. *Eur Radiol* (2021) 32(1):680–9. doi: 10.1038/s41598-021-92341-6

29. Li W, Huang Y, Zhuang BW, Liu GJ, Hu HT, Li X, et al. Multiparametric ultrasomics of significant liver fibrosis: A machine learning-based analysis. *Eur Radiol* (2019) 29(3):1496–506. doi: 10.1007/s00330-018-5680-z

30. Morris DC, Chan DY, Lye TH, Chen H, Palmeri ML, Polascik TJ, et al. Multiparametric ultrasound for targeting prostate cancer: Combining ARFI, SWEI, QUS and b-mode. *Ultrasound Med Biol* (2020) 46(12):3426–39. doi: 10.1016/j.ultrasmedbio.2020.08.022

31. Rosenkrantz AB, Mannelli L, Kong X, Niver BE, Berkman DS, Babb JS, et al. Prostate cancer: utility of fusion of T2-weighted and high b-value diffusion-weighted images for peripheral zone tumor detection and localization. *J Magn Reson Imaging* (2011) 34(1):95–100. doi: 10.1002/jmri.22598

32. Katahira K, Takahara T, Kwee TC, Oda S, Suzuki Y, Morishita S, et al. Ultra-high-b-value diffusion-weighted MR imaging for the detection of prostate cancer: evaluation in 201 cases with histopathological correlation. *Eur Radiol* (2011) 21 (1):188–96. doi: 10.1007/s00330-010-1883-7

frontiers | Frontiers in Oncology

TYPE Systematic Review
PUBLISHED 02 September 2022
DOI 10.3389/fonc.2022.975183

OPEN ACCESS

EDITED BY
Jian Lu,
Peking University Third Hospital, China

REVIEWED BY
Arnaldo Stanzione,
University of Naples Federico II, Italy
Nithesh Naik,
Manipal Academy of Higher Education, India

*CORRESPONDENCE
Jun Pang
pangjun2@mail.sysu.edu.cn

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 22 June 2022
ACCEPTED 17 August 2022
PUBLISHED 02 September 2022

CITATION
Zhang H, Lei H and Pang J (2022) Diagnostic performance of radiomics in adrenal masses: A systematic review and meta-analysis.
Front. Oncol. 12:975183.
doi: 10.3389/fonc.2022.975183

Diagnostic performance of radiomics in adrenal masses: A systematic review and meta-analysis

Hao Zhang, Hanqi Lei and Jun Pang*

Department of Urology, Kidney and Urology Center, Pelvic Floor Disorders Center, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, China

Objectives: (1) To assess the methodological quality and risk of bias of radiomics studies investigating the diagnostic performance in adrenal masses and (2) to determine the potential diagnostic value of radiomics in adrenal tumors by quantitative analysis.

Methods: PubMed, Embase, Web of Science, and Cochrane Library databases were searched for eligible literature. Methodological quality and risk of bias in the included studies were assessed by the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) and Radiomics Quality Score (RQS). The diagnostic performance was evaluated by pooled sensitivity, specificity, diagnostic odds ratio (DOR), and area under the curve (AUC). Spearman's correlation coefficient and subgroup analysis were used to investigate the cause of heterogeneity. Publication bias was examined using the Deeks' funnel plot.

Results: Twenty-eight studies investigating the diagnostic performance of radiomics in adrenal tumors were identified, with a total of 3579 samples. The average RQS was 5.11 (14.2% of total) with an acceptable inter-rater agreement (ICC 0.94, 95% CI 0.93–0.95). The risk of bias was moderate according to the result of QUADAS-2. Nine studies investigating the use of CT-based radiomics in differentiating malignant from benign adrenal tumors were included in the quantitative analysis. The pooled sensitivity, specificity, DOR and AUC with 95% confidence intervals were 0.80 (0.68-0.88), 0.83 (0.73-0.90), 19.06 (7.87-46.19) and 0.88 (0.85–0.91), respectively. There was significant heterogeneity among the included studies but no threshold effect in the meta-analysis. The result of subgroup analysis demonstrated that radiomics based on unenhanced and contrast-enhanced CT possessed higher diagnostic performance, and second-order or higher-order features could enhance the diagnostic sensitivity but also increase the false positive rate. No significant difference in diagnostic ability was observed between studies with machine learning and those without.

Conclusions: The methodological quality and risk of bias of studies investigating the diagnostic performance of radiomics in adrenal tumors should be further improved in the future. CT-based radiomics has the

potential benefits in differentiating malignant from benign adrenal tumors. The heterogeneity between the included studies was a major limitation to obtaining more accurate conclusions.

Systematic Review Registration: https://www.crd.york.ac.uk/PROSPERO/ CRD 42022331999 .

KEYWORDS

adrenal tumor, radiomics, machine learning, diagnostic performance, radiomics quality score

Introduction

Due to the increasing use of abdominal imaging, the discovery of adrenal incidentalomas has kept rising. It is reported that adrenal incidentalomas account for 4-5% of patients without malignancy (1). Although most adrenal masses are benign and non-functional, their functional status and malignant potential should be evaluated when they are detected, according to the latest recommendations (2). However, it is challenging for radiologists to accurately diagnose adrenal masses *via* conventional imaging assessments (3, 4). To begin, imaging features of pitfalls and mimics that are related to various abnormalities and aberrant appearances may potentially lead to misdiagnosis (3). For example, large adenomas usually present as heterogeneous masses on computed tomography (CT) images can not be easily differentiated from adrenocortical carcinoma visually (5, 6). Secondly, conventional imaging assessments depend largely on the experience and knowledge level of the radiologist. Consequently, exploring better approaches to improve the diagnostic value of adrenal imaging is crucial, considering that inappropriate diagnosis can lead to increased treatment costs or unnecessary examination (7).

Radiomics, first pioneered by Philippe Lambin, generally aims to extract quantitative and reproducible data that are imperceptible to the human eye from biomedical images for a series of medical purposes (8, 9). Extracted features, divided into shape-based, first-, second-, and higher-order statistics, can be translated into high-throughput and quantitative data for analysis (10, 11). The features that contribute the most to the objective will be selected for constructing the model *via* statistical approaches and artificial intelligence. Furthermore, radiomics features may achieve complementarity and improve accuracy when combined with clinically acquired, treatment-related, and genomic data (12). As artificial intelligence advances by leaps and bounds, radiomics has been extensively tested and applied in various aspects of oncology, including diagnosis, classification, and prognosis prediction (10). Recently, an increasing number of studies also established that radiomics could offer a risk-free and efficient method to increase the value of diagnostic imaging of adrenal masses. Nakajo et al. investigated the diagnostic performance of standardized uptake value (SUV)-related and texture parameters of F-18-fluorodeoxyglucose positron emission tomography/ computedtomography (FDG PET/CT) between benign and metastatic adrenal tumors (13). In one study, texture analysis was applied to evaluate CT-abnormal adrenal glands in order to differentiate between malignant and benign tumors in patients with lung cancer (14). Moreover, Kong et al. designed a radiomic-based nomogram for pheochromocytoma diagnosis and achieved robust performance (15).

Although radiomics offers a relatively objective and quantitative diagnostic pattern, it is also subjected to data collection, radiomics characteristics processing, and modeling methods. Considering that the quality and results of published studies are mixed, diagnostic performance and feasibility of radiomics in adrenal masses remain elusive. Hence, the aim of the present review was to assess the methodological quality and risk of bias of radiomics studies investigating diagnostic performance in adrenal masses and to determine the potential diagnostic value of radiomics in adrenal tumors by quantitative analysis.

Materials and methods

This review followed the Cochrane Handbook for Systematic Reviews of Interventions and was conducted in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy) statement (16, 17). The protocol of this review is available through PROSPERO (CRD 42022331999).

Literature search

PubMed, Embase, Web of Science, and Cochrane Library databases were searched by two independent observers to identify eligible studies in May 2022. Additionally, the

reference lists of the included studies were manually searched for studies that might meet the inclusion criteria.

Study selection

The titles and abstracts of potentially relevant studies were screened by two reviewers (HZ and HL) independently. Then, the same two reviewers analyzed the full texts of eligible studies and determined the pieces of literature that met the inclusion criteria. Discrepancies between the two investigators were resolved by consensus with a third reviewer (JP).

All single, comparative studies, and primary studies that met the following PICO criteria were selected:

P (patients): Patients with benign or malign adrenal tumors;

I (interventions): Radiomics or texture analysis;

C (comparison): Standard-of-care imaging including computed tomography (CT) and magnetic resonance imaging (MRI), and positron emission tomography/computedtomography (PET/CT);

O (outcome): Histologic typing (including differentiation between different adrenal masses and differentiation between benign and malign adrenal tumors).

The exclusion criteria were as follows: (a) letters, reviews, editorials, expert opinions, case reports, meeting abstracts and comments; (b) non-human research; (c) the study was not written in English. The full search terms are outlined in Table S1.

Quality assessment

The Radiomics Quality Score (RQS) and Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tools were utilized to assess the methodological quality and risk of bias of the included studies, respectively (18, 19). RQS comprises a total of 16 criteria, and the score of each item corresponds to the importance of the methodological quality of the study. The total score ranges from -8 to +36 points, with -8 to 0 points defined as 0% and 36 as 100% (16). The QUADAS-2 tool includes four evaluation criteria: (a) patient selection; (b) index test; (c) reference standard; and (d) flow and timing. Two independent reviewers (HZ and HL) performed the quality assessment, and disagreements between the two reviewers were resolved by consensus with a third reviewer (JP).

Meta-analysis

A meta-analysis may be performed only when a sufficient number of studies attempt to answer a similar question. In this study, we performed a meta-analysis of all studies investigating the diagnostic performance of CT-based radiomics between malign and benign adrenal tumors. Data from all the eligible studies were extracted by two independent reviewers (HZ and HL). Discrepancies were resolved by consensus with a third reviewer (JP). Only studies from which a two-by-two contingency table could be extracted or reconstructed were included. If there were multiple models in the study, only the one with the highest area under the curve (AUC) was extracted. Moreover, the data from the model with the highest Youden's Index was selected if AUC was not reported. When multiple publications were from the same research, only the study with the higher methodological quality was included.

Statistical analysis

Pooled sensitivity, specificity, diagnostic odds ratio (DOR), positive likelihood ratio (PLR), and negative likelihood ratio (NLR) with 95% confidence intervals (CIs) were employed to quantify the diagnostic performance. In addition, diagnostic accuracy was outlined by the summary receiver operating characteristic curve (SROC) and area under the curve (AUC). The heterogeneity of studies was assessed by calculating the I^2 index, where an I^2 value of 0–25% represents insignificant heterogeneity, >25–50% indicates low heterogeneity, >50–75% indicates moderate heterogeneity, and >75% indicates high heterogeneity (20). A $p < 0.05$ was considered statistically significant. A random-effects model was employed to evaluate effect size and pool studies. Forest plots were constructed for visualization of the results. Spearman's correlation coefficient was used to assess the threshold effect between sensitivity logit and (1-specificity) logit. In order to investigate the source of heterogeneity, a subgroup analysis was also conducted with the following covariates: (a) CT Type; (b) CT Feature Type; (c) Machine Learning; (d) Reference. The sensitivity analysis was performed by eliminating the included studies one after another. Publication bias was explored using the Deeks' funnel plot, and statistical significance was assessed by Deeks' asymmetry test. Clinical utility was examined using a Fagan plot, which provided the posttest probability when pretest probabilities were calculated (21).

Stata software (Stata Corporation, College Station, TX, USA, version 16.0) and the Open Meta-analyst (a completely open-source, cross-platform software) were used to conduct the meta-analysis. The interclass correlation coefficient (ICC), which described inter-rater agreement for the RQS and QUADAS-2 and Spearman's correlation coefficient, was determined by SPSS software (IBM, Armonk, NY, USA, version 25.0).

Results

Included studies

The PRISMA flow-chart of the literature search of this systematic review and meta-analysis is presented in Figure 1.

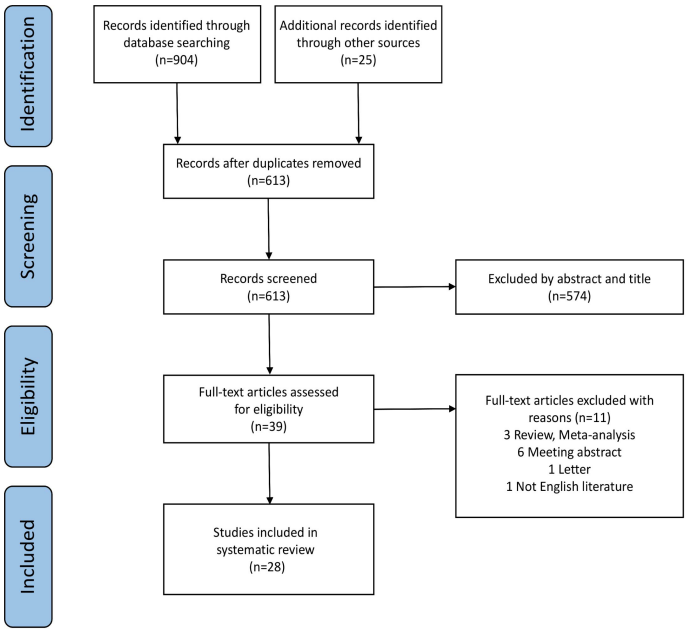


FIGURE 1
Flow diagram of study selection.

613 studies were screened following the removal of 316 duplicate records. Then, 574 articles were excluded by evaluating the abstract and title. After thoroughly screening the full-text, 3 studies were excluded for being reviews or meta-analyses; 6 for being meeting abstracts; one for being a letter; and one for being in a non-English language. Eventually, 28 studies were enrolled in this research. Table 1 summarizes the characteristics of the included studies.

All 28 studies were retrospective cohort studies, and the sample size (number of lesions) ranged from 19 to 377. Most objectives of the included studies were differentiation between benign and malignant adrenal neoplasms by radiomics, followed by the differentiation between pheochromocytoma and adenoma. Other studies distinguished adrenal adenomas from non-adenomas or identified subtypes of adrenal adenomas. The majority of studies focused on the diagnostic performance of radiomics using CT imaging (n=19), while a quarter was based on magnetic resonance imaging (n=7). Additionally, one study explored the use of radiomics based on a combination of CT and MR and one based on FDG PET/CT. More than 78% (22/28) of studies used manual segmentation. Radiomics feature types used by different studies varied. Interestingly, over half of the included studies (n=17) extracted second or higher-order features for analyses. As for the modeling method, 12 studies conducted logistic regression, eight studies did not provide relevant information and the remainder employed other algorithms such as support vector machine, random forest, extra trees classifier and so forth. More than half studies utilized histopathology as the gold reference (n=15). Ten studies combined histopathology and follow-up imaging. Two articles exclusively considered clinical and imaging follow-ups, and one study failed to report the reference standard. Outcomes of the included studies are summarized in Table 2.

Data quality assessment

The included studies achieved an mean ± standard deviation RQS of 5.11 ± 7.70, a median of 3.5, interquartile range 14, and a range of -5 to 25. The mean RQS proportion was 14.2%, with a maximum of 69.4%. The mode scores for the 16 dimensions are summarized in Table 3. The individual scores of each study and final scores of RQS are presented in Tables S2 and S3, respectively.

The majority of studies provided details about the imaging scheme, applied discrimination statistics and achieved their potential clinical utility. Conversely, none of the included studies employed phantoms, considered biological correlates or assessed the repeatability of radiomics analysis at multiple time points. Moreover, feature reduction or adjustment of multiple tests were performed in 16/28 (57%) studies, and non-radiomics features were applied in 3/28 (11%) studies. Only a few studies conducted model calibration, assessed the cost-effectiveness and publicly shared segmentations or code.

TABLE 1 Characteristics of the included studies.

Study ID	Ref	Study Design	Diagnostic Subject	Sample Size	Image Modality	Segmentation Method (Software/Algorithm)	Feature Extraction	Features Type	Modeling method	Reference Standard	Validation
Andersen et al. (2021)	(14)	Retrospective	Adrenal metastases/Benign lesions	160	Contrast-enhanced CT	Semi-automatic (Philips Intellispace Tumor Tracking)	TexRAD	First-order and higher-order	Binary logistic regression	Histopathology	NR
Chai et al. (2017)	(22)	Retrospective	Aldosterone-producing adenomas/ Pheochromocytomas/Cushing adenomas	218	Unenhanced and contrast-enhanced CT	Automatic (Multiscale sparse representations)	NR	Shape-based, first-order and second-order	Support vector machine, radial basis function network (ML)	Histopathology	Internal validation
Elmohr et al. (2019)	(23)	Retrospective	Adrenocortical carcinomas/Adrenocortical adenomas	54	Unenhanced and contrast-enhanced CT	Manual (Amira Software)	PyRadiomics	Shape-based, first-order and second-order	Logistic regression, boruta random forest	Histopathology	Internal validation
Ho et al. (2019)	(24)	Retrospective	Adrenal malignancy/Lipid-poor adenoma	23	Unenhanced and contrast-enhanced CT, MRI 3T or 1,5T T1 in- and opposed-phase	Manual (Seg3D)	Lesion Tool (software developed by the authors)	Shape-based, first-order and second-order	Logistic regression	Histopathology	NR
Kong et al. (2022)	(15)	Retrospective	Pheochromocytoma/Other adrenal lesions	309	MRI 3T T2w	Semi-automatic (3D Slicer)	3D Slicer	Shape-based, first-order, second-order and higher-order	Logistic regression	Histopathology	Internal and external validation
Koyuncu et al. (2019)	(25)	Retrospective	Adrenal malignant/Benign lesions	114	Contrast-enhanced CT	Semi-automatic (AbSeg)	MATLAB	Shape-based, first-order, second-order and higher-order	Bounded particle swarm optimisation-neural network	NR	Internal validation
Li et al. (2018)	(26)	Retrospective	Adrenal malignant/Benign lesions	210	Unenhanced and contrast-enhanced CT	Manual (NR)	NR	Second-order	Bayesian spatial gaussian process classifier	Histopathology	NR
Liu et al. (2021)	(27)	Retrospective	Adrenal Adenoma/Pheochromocytoma	60	MRI 3T T1 in- and opposed-phase, T2w	Manual (Mazda)	MaZda	First-order	Support vector machine (ML)	Histopathology	Internal validation
Nakajo et al. (2017)	(13)	Retrospective	Adrenal metastases/Benign lesions	35	FDG PET/CT	Semi-automatic (Advantage Windows Workstation)	Python	First-order	NR	Clinical and imaging follow-ups	NR
Moawad et al. (2021)	(28)	Retrospective	Adrenal malignant/Benign lesions	40	Unenhanced and contrast-enhanced CT	Manual (Amira Software)	PyRadiomics	Shape-based, first-order, second-order and higher-order	Random forest (ML)	Histopathology	Internal validation
Rocha et al. (2018)	(29)	Retrospective	Adrenal adenomas/Malignant lesions	108	Unenhanced CT	Manual (OsiriX Software)	OsiriX	First-order	NR	Histopathology or follow-up imaging	NR
Romeo et al. (2018)	(30)	Retrospective	Lipid-rich/Lipid-poor/Nonadenoma adrenal lesions	60	MRI 3T T1w, T2w	Manual (ITK-SNAP)	3D Slicer	First-order and second-order	J48 classifier, Weka software (ML)	Histopathology	Internal validation
Schieda et al. (2017)	(31)	Retrospective	Adrenal metastases/Adrenal adenoma	44	MRI 1.5T or 3T T1 in- and opposed-phase, T2w, GRE	Manual (Image J)	Image J	First-order	Logistic regression	Histopathology or follow-up imaging	NR
Shi et al. (2019)	(32)	Retrospective	Adrenal metastases/Benign lesions	265	Unenhanced and contrast-enhanced CT	Manual (NR)	TexRAD	First-order and higher-order	Logistic regression, support vector machin (ML)	Histopathology or follow-up imaging	Internal validation
Shoemaker et al. (2018)	(33)	Retrospective	Adrenal malignant/Benign lesions; Adrenal functioning/Non-functioning lesions	377	Unenhanced CT	NR	NR	First-order and second-order	Logistic regression	Histopathology	Internal validation
Stanzione et al. (2021)	(34)	Retrospective	Adrenal malignant/Benign lesions	55	MRI 3T T1w, T2w	Manual (ITK-SNAP)	PyRadiomics	Shape-based, first-order, second-order and higher-order	Extra trees classifier (ML)	Histopathology or follow-up imaging	Internal validation

(Continued)

10.3389/fonc.2022.975183

TABLE 1 Continued

Study ID	Ref	Study Design	Diagnostic Subject	Sample Size	Image Modality	Segmentation Method (Software/Algorithm)	Feature Extraction	Features Type	Modeling method	Reference Standard	Validation
Szász et al. (2020)	(35)	Retrospective	Adrenal adenomas/Non-adenomas	233	Unenhanced CT	Manual (Advantage Windows workstation)	"Volume Histogram" tool	First-order	NR	Histopathology	NR
Torresan et al. (2021)	(36)	Retrospective	Adrenocortical carcinomas/Adenoma	19	Unenhanced and contrast-enhanced CT	Manual (PMOD)	PMOD	First-order and second-order	K-means clustering technique (ML)	Histopathology or follow-up imaging	NR
Tu et al. (2018)	(37)	Retrospective	Adrenal metastases/Adenomas	76	Contrast-enhanced CT	Manual (ImageJ)	Image J	First-order	Logistic regression	Previously described imaging thresholds or follow-up imaging	NR
Tu et al. (2020)	(38)	Retrospective	Adrenal Metastases/Lipid-poor adenomas	63	MRI 1.5T or 3T T1w, T2w, GRE	Manual (Image J)	Image J	First-order	Logistic regression	Histopathology or follow-up imaging	Internal validation
Tüdös et al. (2019)	(39)	Retrospective	Adrenal lipid-poor adenomas/Non-adenomas	163	Unenhanced CT	Manual (Advantage Windows workstation)	"Volume Histogram" tool	First-order	NR	Histopathology or follow-up imaging	NR
Umanodan et al. (2017)	(40)	Retrospective	Pheochromocytomas/Adrenal adenomas	52	MRI 3T ADC	Manual (Synapse Vincent software)	Synapse Vincent software	First-order	NR	Histopathology or follow-up imaging	NR
Wu et al. (2020)	(41)	Retrospective	Adrenal adenoma/Nonadenoma	94	Unenhanced CT	Manual (PACS software)	PACS software	First-order	NR	Histopathology or follow-up imaging	NR
Yi et al. (2018)	(42)	Retrospective	Pheochromocytomas/Adrenal lipid-poor adenomas	110	Unenhanced CT	Manual (MaZda)	MaZda	First-order, second-order and higher-order	Logistic regression (ML)	Histopathology	NR
Yi et al. (2018) (2)	(43)	Retrospective	Pheochromocytoma/Adrenal lipid-poor adenoma	265	Unenhanced and contrast-enhanced CT	Manual (MaZda)	MaZda	First-order, second-order and higher-order	Lasso, logistic regression	Histopathology	Internal validation
Yu et al. (2020)	(44)	Retrospective	Adrenal malignant/Benign lesions	125	Contrast-enhanced CT	Manual (TexRAD)	TexRAD	First-order and higher-order	NR	Histopathology or follow-up imaging	NR
Zhang et al. (2017)	(45)	Retrospective	Pheochromocytomas/Lipid-poor adrenocortical adenoma	164	Unenhanced and contrast-enhanced CT	Manual (TexRAD)	TexRAD	First-order and higher-order	NR	Histopathology	NR
Zheng et al. (2020)	(46)	Retrospective	Aldosterone-producing/Cortisol-producing functional adrenocortical adenomas	83	Unenhanced and contrast-enhanced CT	Manual (ITK-SNAP)	NR	Shape-based, first-order	Lasso, logistic regression (ML)	Histopathology	Internal validation

Ref, reference; NR, not report; ML, machine learning; PACS, picture archiving and communication system.

10.3389/fonc.2022.975183

10.3389/fonc.2022.975183

TABLE 2 Outcomes of the included studies.

Study ID	P	N	TP	FP	TN	FN	Sensitivity, %	Specificity, %	Accuracy, %	AUC	95%CI
Andersen et al. (2021)	71	89	41	20	69	30	58	77	68	0.730	–
Chai et al. (2017)	–	–	–	–	–	–	–	–	81.8 ~ 95.4	–	–
Elmohr et al. (2019)	–	–	–	–	–	–	81	83	82	0.890	–
Ho et al. (2019)	8	15	–	–	–	–	–	–	80	–	–
Kong et al. (2022)	–	–	–	–	–	–	85.7	75	84	0.906	0.841-0.971
Koyuncu et al. (2019)	12	45	9	8	37	3	75	82.2	80.7	0.786	–
Li et al. (2018)	96	114	91	37	77	5	94.8	67.5	80	–	–
Liu et al. (2021)	–	–	–	–	–	–	–	–	85	0.917	–
Nakajo et al. (2017)	22	13	22	2	11	0	100	84.6	94.3	0.970	0.840-0.990
Moawad et al. (2021)	19	21	16	6	15	3	84.2	71.4	77.5	0.850	–
Rocha et al. (2018)	88	20	77	1	19	11	87.5	95	88.9	–	–
Romeo et al. (2018)	–	–	–	–	–	–	–	–	80	–	–
Schieda et al. (2017)	15	29	14	4	25	1	93.3	86.2	88.6	0.970	0.930-1.000
Shi et al. (2019)	101	164	78	37	127	23	77	77	77.4	0.850	0.800-0.890
Shoemaker et al. (2018)	–	–	–	–	–	–	–	–	–	0.780~1.000	–
Stanzione et al. (2021)	18	37	–	–	–	–	–	–	0.91	0.970	0.870-1.000
Szász et al. (2020)	123	110	–	–	–	–	–	–	–	0.919	–
Torresan et al. (2021)	8	10	7	1	9	1	87.5	90	88.9	–	–
Tu et al. (2018)	40	36	19	9	27	21	47.5	75	60.5	0.650	0.520-0.770
Tu et al. (2020)	40	23	30	0	23	10	75	100	84.1	0.950	0.910-0.990
Tüdös et al. (2019)	83	80	44	1	79	39	53	98.8	75.5	–	–
Umanodan et al. (2017)	39	13	37	1	12	2	94.9	92.3	94.2	0.920	–
Wu et al. (2020)	58	36	51	16	20	7	87.9	55.6	75.5	0.740	–
Yi et al. (2018)	29	79	25	2	77	4	86.2	97.5	94.4	0.952	0.897-1.000
Yi et al. (2018) (2)	67	145	64	14	131	3	95.5	90.3	92	0.957	–
Yu et al. (2020)	81	44	66	0	44	15	81	100	88	0.970	0.940-0.990
Zhang et al. (2017)	98	66	78	11	55	20	79.6	83.3	81.1	0.860	0.810-0.910
Zheng et al. (2020)	–	–	–	–	–	–	91.5	92.8	92.2	0.902	0.822-0.982

P,condition positive; N, condition negative; TP, true positive; FP, false positive; TN, true negative; FN, false negative; AUC, area under the receiver operating characteristic; CI, confidence interval.

The inter-reader agreement was found to be moderate to excellent for radiomics features in 39% (11/28) of the included studies. Nevertheless, validation of more than half of the included studies was missing (15/28, 50%). Only one study (28) compared the diagnostic performance of the classifier with an expert radiologist, but no significant differences were noted. In general, the quality of included articles was acceptable, and the assessment of the risk of bias and applicability of the 28 included studies are illustrated in Figure 2. The detail of the individual and final evaluation of the risk of bias and applicability concerns are presented in Tables S4 and S5, respectively.

Inter-rater agreements of RQS and QUADAS-2 were also assessed by the ICC. The ICC for the RQS was 0.94 (95% CI 0.93–0.95). Six criteria of RQS reached a moderate agreement, while ten items achieved substantial or almost perfect agreement (Table 4). The ICC for the QUADAS-2 was 0.96 (95% CI 0.95–0.97). Except for two dimensions reaching a moderate agreement, the others exceeded 75% agreement (Table 5).

Meta-analysis

We performed a meta-analysis investigating the use of CT-based radiomics in differentiating malignant from benign adrenal tumors and enrolled nine eligible studies, from which a two-by-two contingency table could be extracted or reconstructed. As shown in Table 6, the mean values and 95% CIs of the pooled sensitivity, specificity, PLR, NLR, and DOR for the radiomics signature based on CT in differentiating malignant adrenal tumors from benign tumors were 0.80 (0.68-0.88), 0.83 (0.73-0.90), 4.70 (2.80-8.00), 0.25 (0.15-0.41) and 19.06 (7.87-46.19) respectively. The summary receiver operating characteristic curve showed an overall pooled AUC of 0.88 (95% CI 0.85–0.91) (Figure 3). Significant heterogeneity in sensitivity (I^2 = 87.09%) and specificity (I^2 = 72.1%) were noted, as depicted in Figure 4. Consequently, diagnostic threshold analysis was carried out, which revealed that there was no threshold effect, given that the Spearman's correlation coefficient was -0.036 and the *p*-value was 0.932. In order to

10.3389/fonc.2022.975183

TABLE 3 Elements of the RQS and average rating achieved by the studies included in this systematic review.

RQS scoring item	Interpretation	Mode
Image Protocol	+1 for well documented protocols, +1 for publicly available protocols	1
Multiple Segmentations	+1 if segmented multiple times (different physicians, algorithms, or perturbation of regions of interest)	1
Phantom Study	+1 if texture phantoms were used for feature robustness assessment	0
Multiple Time Points	+1 multiple time points for feature robustness assessment	0
Feature Reduction	−3 if nothing, +3 if either feature reduction or correction for multiple testing	3
Non Radiomics	+1 if multivariable analysis with non-radiomics features	0
Biological Correlates	+1 if present	0
Cut-off	+1 if cutoff either pre-defined or at median or continuous risk variable reported	0
Discrimination and Resampling	+1 for discrimination statistic and statistical significance, +1 if resampling applied	1
Calibration	+1 for calibration statistic and statistical significance, +1 if resampling applied	0
Prospective	+7 for prospective validation within a registered study	0
Validation	−5 if no validation/+2 for internal validation/+3 for external validation/+4 two external validation datasets or validation of previously published signature/+5 validation on ≥3 datasets from >1 institute	-5
Gold Standard	+2 for comparison to gold standard	2
Clinical Utility	+2 for reporting potential clinical utility	2
Cost-effectiveness	+1 for cost-effectiveness analysis	0
Open Science	+1 for open-source scans, +1 for open-source segmentations, +1 for open-source code, +1 open-source representative segmentations and features	0

further explore the cause of heterogeneity, subgroup analysis was also performed, as outlined in Table 6.

Four studies with radiomics based on unenhanced and contrast-enhanced CT demonstrated higher sensitivity (0.87 vs. 0.66) but lower specificity (0.74 vs. 0.80) than studies using only contrast-enhanced CT. Studies (n=2) that only included first-order features had lower sensitivity (0.72 vs. 0.81) but higher specificity (0.86 vs. 0.77) compared to those that combined with second-order or higher-order features. Interestingly, the studies (n=3) that applied machine learning gained equivalent sensitivity (0.79 vs. 0.78) as well as specificity (0.77 vs. 0.79) compared to those did not use (n=6). Four studies that considered histopathology or follow-up imaging as a reference had higher specificity (0.91 vs. 0.72) and equivalent sensitivity (0.82 vs. 0.83) than studies (n=3) using only histopathology. The corresponding forest plots for sensitivity and specificity are delineated in Figures S1–S4.

As shown in Table S6, we can hardly identify significant changes in the pooled effect value when eliminating studies one by one. There was no publication bias based on the Deeks' funnel plot (p=0.77), as presented in Figure 5. Furthermore, the clinical utility was also evaluated using a Fagan plot. Using a CT-based radiomics model would increase the posttest probability to 54% from 20% with a PLR of 5 when the pretest was positive and reduce the posttest probability to 6% with an NLR of 0.25 when the pretest was negative, as depicted in Figure S5.

Discussion

Radiomics has recently attracted the attention of oncology researchers, given that it can noninvasively and effectively reflect tumor heterogeneity, treatment response, prognosis, and other information (47–49). Published studies involving radiomics for

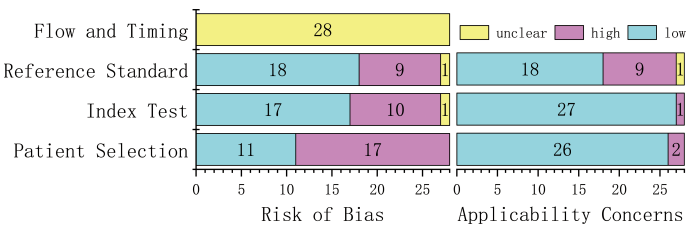


FIGURE 2
The risk of bias and concerns regarding applicability of 28 included studies.

TABLE 4 Inter-rater agreement in RQS assessment.

RQS scoring item	ICC (95% CI)
Image Protocol	0.52 (0.19-0.75)
Multiple Segmentations	0.93 (0.86-0.97)
Phantom Study	1.00 (1.00–1.00)
Multiple Time Points	1.00 (1.00–1.00)
Feature Reduction	0.86 (0.72-0.93)
Non Radiomics	0.79 (0.59-0.90)
Biological Correlates	1.00 (1.00–1.00)
Cut-off	0.63 (0.34-0.81)
Discrimination and Resampling	0.52 (0.19-0.75)
Calibration	1.00 (1.00–1.00)
Prospective	1.00 (1.00–1.00)
Validation	1.00 (1.00–1.00)
Gold Standard	0.54 (0.22–0.76)
Clinical Utility	0.61 (0.32-0.80)
Cost-effectiveness	1.00 (1.00–1.00)
Open Science	0.79 (0.59-0.90)

CI: confidence interval, RQS: Radiomics Quality Score.

adrenal tumors mainly focused on the differentiation of benign and malignant tumors and the differentiation of histological types of adrenal masses, but no clinical transformation or practical application has been described so far. Considering that the research of radiomics remains in its infancy, there are a host of problems to be addressed, such as the lack of a robust workflow based on standardized and strict methods to ensure the stability and reliability of the results (50, 51).

It is well-established that the quality of reporting of existing predictive models is poor (52). Lambin et al. proposed a comprehensive and clear standard to evaluate all aspects of predictive models in the field of radiomics to enhance their qualities (18). In our systematic review, the overall quality was relatively low (mean RQS of 5.11, ranging from -5 to 25). The primary causes impacting the RQS score included the absence of feature reduction, scarcity of open science and source, deficiency in internal or external validation and prospective data support, minimal consideration of cost-effectiveness, and so on. Reducing features that are poorly reproducible is crucial for reducing the risk of overfitting when the number of radiomics features exceeds the number of patients (53). 43% of studies did not conduct feature reduction or adjustment in our review owing to specific diagnostic algorithms or processing, which may partially undermine the stability of the models. Following internal and external validation, the diagnostic performance of the model can be confirmed. Furthermore, the practical application of radiomics in clinical practice also requires multi-center validation and prospective testing (54, 55). Regrettably, over half of the included studies failed to process validation attributable to the limited sample size. Most of the remaining studies merely conducted internal validation in a single center. Furthermore, only one study performed validation with multiple data sets and tested prospectively (15). Furthermore, comparing the diagnostic performance with the radiologist is also a pivotal step in verifying the performance of the model. Because only when the diagnostic effect is better than that of the radiologist can the superiority of radiomics be demonstrated. However, most studies did not compare the diagnostic performance with a radiologist. The choice of scanner manufacturer and model, 2D or 3D segmentation of the region of interest, acquisition, and reconstruction parameters all lead to the heterogeneity of imaging data. Most studies (25/28) provided image acquisition parameters in our review, but values varied considerably. Zwanenburg et al. designed the Image Biomarker Standardization Initiative (IBSI) to enhance the reproducibility of radiomics research, including establishing general feature naming, definition, general radiological image processing scheme, and so on (56). Thus, open science and the source of radiomics is the premise to realizing reproducibility. In the present review, only three included studies publicly shared segmentations or code. The challenge of open science and validation based on a sufficient sample size may hinder further development and practice of radiomics in the diagnosis of adrenal masses. Besides, the cost-effectiveness analyses of radiomics cannot be overlooked because it may boost the superiority of this technology.

TABLE 5 Inter-rater agreement in QUADAS-2 assessment.

RQS scoring item	ICC (95% CI)
Risk of Bias - Patient Selection	0.79 (0.60-0.90)
Risk of Bias - Index Test	0.94 (0.87-0.97)
Risk of Bias - Reference Standard	1.00 (1.00–1.00)
Risk of Bias - Flow and Timing	1.00 (1.00–1.00)
Applicability Concerns- Patient Selection	0.52 (0.19-0.75)
Applicability Concerns- Index Test	0.66 (0.39-0.83)
Applicability Concerns- Reference Standard	1.00 (1.00–1.00)

CI: confidence interval, QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2.

Although radiomics studies differ methodologically from conventional trials, and there may be potential unsuitability of the QUADAS-2 tool, the results that reflected the risk of bias and applicability of included studies is advisable to some extent. The results of QUADAS-2 exposed that the risk of bias needs to be minimized in terms of patient selection, index test, and reference standards. The concerns regarding applicability are excellent except for the reference standard. The reliablity of individual ratings needs to be assessed by inter-rater agreement analysis. In this review, the ICC was applied to describe the inter-rater agreement of RQS and QUADAS-2. The fact that most items achieved substantial or almost perfect agreement while others had moderate agreement demonstrates that the scores accurately reflect the quality of the included studies.

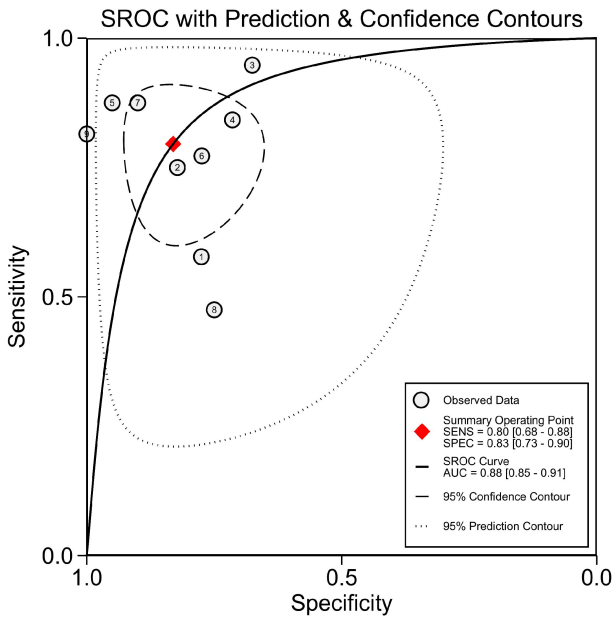


FIGURE 3
Summary receiver operating characteristic (SROC). AUC, area under the curve.

Since there are few pieces of literature included in the meta-analysis, the results should be treated with caution. In our meta-analysis, radiomics technology showed promise for differentiating malignant from benign adrenal tumors, with a pooled sensitivity, specificity, and AUC of 0.8, 0.83, and 0.88, respectively. Nonetheless, it cannot be ignored that there was distinct heterogeneity between the studies. The threshold effect is one of the chief causes of heterogeneity in DTA studies (57). A threshold effect will result in a correlation coefficient between sensitivity and a false positive rate of 0.6 or higher (58). The result of the Spearman correlation coefficient showed no threshold effect in this meta-analysis. Consequently, we attempted to determine the causes of heterogeneity *via* subgroup analysis. Our results demonstrated that the radiomics group based on unenhanced and contrast-enhanced CT had a higher DOR than studies using contrast-enhanced CT

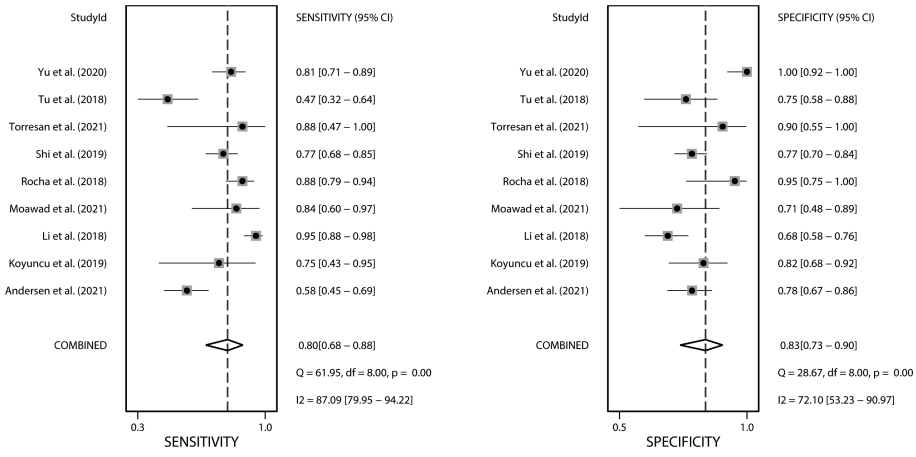


FIGURE 4
Forest plots of the sensitivity and specificity of CT-based radiomics in differentiating malignant from benign adrenal tumors. I^2 >50% indicates substantial heterogeneity among included studies.

10.3389/fonc.2022.975183

TABLE 6 The results of subgroup analysis.

Analysis	No. of study	Sensitivity	Specificity	PLR	NLR	DOR
CT Type						
Contrast-enhanced CT	4	0.66(0.47-0.80)	0.80(0.70-0.88)	3.15(1.69-5.89)	0.50(0.35-0.72)	9.02(2.59-31.43)
Unenhanced and contrast-enhanced CT	4	0.87(0.72-0.95)	0.74(0.66-0.80)	3.15(2.60-3.82)	0.17(0.07-0.41)	18.89(8.96-39.85)
Unenhanced CT	1	0.88(0.79-0.93)	0.95(0.72-0.99)	17.50(2.59-118.41)	0.13(0.02-0.89)	133.00(16.16-1094.59)
CT Feature Type						
With second-order or higher-order features	7	0.81(0.69-0.89)	0.77(0.70-0.83)	3.21(2.55-4.04)	0.23(0.11-0.47)	16.97(7.56-38.12)
Only first-order	2	0.72(0.25-0.95)	0.86(0.51-0.97)	4.77(0.56-40.72)	0.39(0.08-1.86)	16.91(0.38-761.14)
Machine Learing						
Not use machine learing	6	0.78(0.60-0.89)	0.79(0.69-0.87)	3.20(2.12-4.82)	0.26(0.09-0.76)	18.80(5.37-65.75)
Use machine learing	3	0.79(0.71-0.85)	0.77(0.71-0.83)	3.41(2.59-4.50)	0.28(0.21-0.37)	12.54(7.28-21.59)
Reference						
Histopathology	3	0.83(0.46-0.97)	0.72(0.65-0.78)	2.83(2.28-3.52)	0.21(0.05-0.81)	12.93(3.08-54.26)
Histopathology or follow-up imaging	4	0.82(0.76-0.86)	0.91(0.72-0.98)	9.33(2.60-33.52)	0.28(0.21-0.38)	59.05(9.39-371.52)
Previously described imaging thresholds or follow-up imaging	1	0.48(0.33-0.63)	0.75(0.59-0.86)	1.90(0.99-3.65)	0.70(0.36-1.35)	2.71(1.02-7.21)
Overall	9	0.80(0.68-0.88)	0.83(0.73-0.90)	4.70(2.80-8.00)	0.25(0.15-0.41)	19.06(7.87-46.19)

PLR, positive likelihood ratio; NLR, negative likelihood ratio; DOR, diagnostic odds ratio; The 95% confidence intervals are shown in parentheses.

only. This is likely due to the fact that unenhanced CT provides additional features for analysis and bring higher sensitivity. Different levels of radiomics features contain distinct dimensions of information regarding the lesion. First-order statistics features describe the distribution of voxel values without concern for their spatial relationships (11). Second-order statistics, which describe spatial relationships between voxels with similar gray levels within a lesion, can provide a measure of intralesional heterogeneity (11, 59). Higher-order statistics are obtained after imposing filter grids on an image, and the processing can confirm repetitive or non-repetitive patterns, suppress image noise, highlight details, and so on (60). According to our results, studies that only included first-order features had lower sensitivity but higher specificity compared to those that combined second-order or higher-order features. This finding signals that more complex and deeper texture features analyses can improve diagnostic sensitivity while also increasing the false positive rate. Since deeper texture features analyses inevitably yield a large number of unstable and unrepeatable features, advanced features have higher requirements for feature selection and modeling algorithms. Machine learning is a broad term for a class of statistical analysis algorithms that can iteratively improve the predictive performance of a model by "learning" from data (61). Reliable machine-learning approaches can drive the success of radiomic applications in clinical care (62). In our subgroup analysis, studies with machine learning achieved equivalent diagnostic performance to those without. However, the

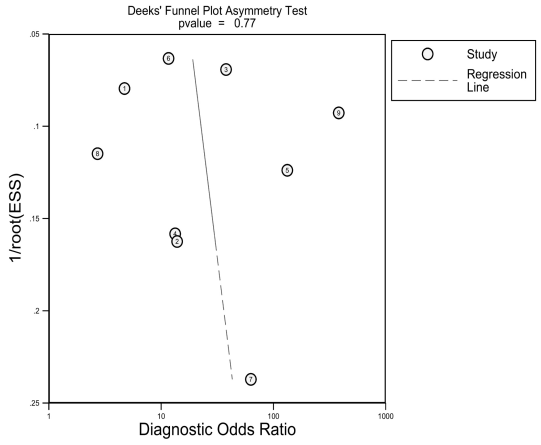


FIGURE 5
Deeks funnel plot reveals the possibility of publication bias is low with a p value of 0.77. ESS, effective sample size.

number of studies (n=3) is insufficient to represent the true impact of machine learning. We recommend that further studies be conducted to determine whether machine learning is beneficial to the diagnosis of adrenal tumors using radiomics. Some studies included in this meta-analysis enrolled patients without histopathology results and regarded follow-up imaging as the diagnostic reference (29, 32, 36, 37, 44). On the one hand, these studies may lower selection bias, as potential bias will be generated if studies only include patients who underwent surgeries (those with high suspicion of malignancy are more likely to be operated on). On the other hand, the diagnostic accuracy of this method based on follow-up images remains to be determined. The 2017 American College of Radiology white paper (63) suggests that stability for 1 year or more indicates that uncertain adrenal nodules are benign, whereas enlarged nodules are suspected to be malignant. However, benign tumors can also grow, and the threshold growth rate to consider malignancy remains unknown. Studies that regarded histopathology or follow-up imaging as the reference had higher diagnostic specificity than studies using histopathology only. The reason may be that the true negative ratio was overestimated since some follow-up imaging failed to detect potential malignancies. Additionally, the possibility that heterogeneity was caused by other factors that have not been considered cannot be ruled out.

To the best of our knowledge, there are two previous reviews related to similar topics. One study systematically reviewed the diagnostic accuracy of CT texture analysis in adrenal tumors (64). In another review, Stanzione et al. summarized the application of radiomics in adrenal cross-sectional imaging and assessed the methodological quality by RQS (65). Generally, more comprehensive and in-depth analyses of diagnostic performance of radiomics in adrenal masses were done in our study. First of all, we focused on diagnostic performance of radiomics in various radiological imaging of adrenal tumors. Secondly, RQS and QUADAS-2 of the included studies were independently evaluated by two reviewers. Besides, inter-rater agreement for RQS and QUADAS-2 were also assessed, which can reflect the true quality of the included studies better. In addition to a systematic review of the included studies, we also conducted a meta-analysis investigating the role of CT-based radiomics in differentiating malignant from benign adrenal tumors. Although the heterogeneity was significant, it reflected the diagnostic value of radiomics in differentiating benign and malignant adrenal masses to some extent.

There are several limitations of this review that warrant consideration. To begin, grey literature was not included in this review since it was limited to special circulation channels, which might have led to publication bias. Secondly, the overall quality of the included studies was not optimal (mean RQS 14.2%), which may have partly influenced the quality of the subsequent analysis. Thirdly, it is worthwhile mentioning the heterogeneity of studies included in the quantitative synthesis. Except for CT type, CT feature type, machine learning, and diagnostic reference, the heterogeneity may be pertinent to diversity in pathological types, methods of image segmentation and reconstruction, and feature extraction and modeling algorithms. However, because the subgroup distribution was scattered, we were unable to analyze these detailed features. Hence, the results of the quantitative analysis should be interpreted with caution. Fourthly, the diagnostic performance of radiomics between specific adrenal histologic types could not be assessed because of a lack of studies for the same objective. Lastly, given that only a few studies compared the diagnostic performance with a radiologist, the added value of radiomics in comparison to the accuracy of human assessment could not be explored.

Conclusion

In conclusion, we systematically reviewed studies investigating the diagnostic performance of radiomics in adrenal masses and conducted a meta-analysis. Collectively, the results of quantitative synthesis outline the potential benefits of CT-based radiomics in differentiating malignant from benign adrenal tumors. However, the existing limitations of relevant studies, including the lack of validation and prospective tests, the lack of comparison with a radiologist, and the absence of a standardized radiomics process, hinder the further development of radiomics. We postulate that the translational gap between radiomics research and clinical applications in the field of adrenal tumors diagnosis will be overcome in the future by addressing the aforementioned shortcomings.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization and methodology: HZ and HL. Data analysis, drafting and revising: HZ, HL, and JP. Writing, editing, and revision of manuscript: HZ. Supervision, review, project administration and funding acquisition: JP. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Sanming Project of Medicine in Shenzhen (00101100032).

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.975183/full#supplementary-material

References

1. Song JH, Chaudhry FS, Mayo-Smith WW. The incidental adrenal mass on CT: Prevalence of adrenal disease in 1,049 consecutive adrenal masses in patients with no known malignancy. *Am J Roentgenol* (2008) 190:1163–8. doi: 10.2214/AJR.07.2799

2. Beltsevich D, Troshina E, Melnichenko G, Platonova N, Ladygina D, Chevais A. Draft of the clinical practice guidelines "Adrenal incidentaloma". *Endo Serg* (2022) 15(1):4–26. doi: 10.14341/serg12712

3. Elsayes KM, Elmohr MM, Javadi S, Menias CO, Remer EM, Morani AC, et al. Mimics, pitfalls, and misdiagnoses of adrenal masses on CT and MRI. *Abdom Radiol* (2020) 45:982–1000. doi: 10.1007/s00261-019-02082-4

4. Anagnostis P, Karagiannis A, Tziomalos K, Kakafika AI, Athyros VG, Mikhailidis DP. Adrenal incidentaloma: a diagnostic challenge. *Hormones (Athens)* (2009) 8:163–84. doi: 10.14310/horm.2002.1233

5. Johnson PT, Horton KM, Fishman EK. Adrenal mass imaging with multidetector CT: pathologic conditions, pearls, and pitfalls. *Radiographics* (2009) 29:1333–51. doi: 10.1148/rg.295095027

6. Park SY, Park BK, Park JJ, Kim CK. CT sensitivities for large (≥3 cm) adrenal adenoma and cortical carcinoma. *Abdom Imaging* (2015) 40:310–7. doi: 10.1007/s00261-014-0202-1

7. Pandharipande PV, Herts BR, Gore RM, Mayo-Smith WW, Harvey HB, Megibow AJ, et al. Rethinking normal: Benefits and risks of not reporting harmless incidental findings. *J Am Coll Radiol* (2016) 13:764–7. doi: 10.1016/j.jacr.2016.03.017

8. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology* (2016) 278:563–77. doi: 10.1148/radiol.2015151169

9. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036

10. Ding H, Wu C, Liao N, Zhan Q, Sun W, Huang Y, et al. Radiomics in oncology: A 10-year bibliometric analysis. *Front Oncol* (2021) 11:689802. doi: 10.3389/fonc.2021.689802

11. Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi AM, Mirabelli P, et al. Radiogenomic analysis of oncological data: A technical survey. *Int J Mol Sci* (2017) 18:1–28. doi: 10.3390/ijms18040805

12. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. *Radiology* (2013) 269:8–15. doi: 10.1148/radiol.13122697

13. Nakajo M, Jinguji M, Nakajo M, Shinaji T, Nakabeppu Y, Fukukura Y, et al. Texture analysis of FDG PET/CT for differentiating between FDG-avid benign and metastatic adrenal tumors: efficacy of combining SUV and texture parameters. *Abdom Radiol (New York)* (2017) 42:2882–9. doi: 10.1007/s00261-017-1207-3

14. Andersen MB, Bodtger U, Andersen IR, Thorup KS, Ganeshan B, Rasmussen F. Metastases or benign adrenal lesions in patients with histopathological verification of lung cancer: Can CT texture analysis distinguish? *Eur J Radiol* (2021) 138:109664. doi: 10.1016/j.ejrad.2021.109664

15. Kong J, Zheng J, Wu J, Wu S, Cai J, Diao X, et al. Development of a radiomics model to diagnose pheochromocytoma preoperatively: a multicenter study with prospective validation. *J Transl Med* (2022) 20:31. doi: 10.1186/s12967-022-03233-w

16. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. *Cochrane Database Syst Rev* (2019) 10:ED000142. doi: 10.1002/14651858.ED000142

17. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies the PRISMA-DTA statement. *JAMA - J Am Med Assoc* (2018) 319:388–96. doi: 10.1001/jama.2017.19163

18. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (2017) 14:749–62. doi: 10.1038/nrclinonc.2017.141

19. Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM, Whiting PF, Rutjes AWSS, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* (2011) 155:529–36. doi: 10.7326/0003-4819-155-8-201110180-00009

20. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* (2003) 327:557–60. doi: 10.1136/bmj.327.7414.557

21. Hellmich M, Lehmacher W. A ruler for interpreting diagnostic test results. *Methods Inf Med* (2005) 44:124–6. doi: 10.1055/s-0038-1633930

22. Chai H, Guo Y, Wang Y, Zhou G. Automatic computer aided analysis algorithms and system for adrenal tumors on CT images. *Technol Heal Care* (2017) 25:1105–18. doi: 10.3233/THC-160597

23. Elmohr MM, Fuentes D, Habra MA, Bhosale PR, Qayyum AA, Gates E, et al. Machine learning-based texture analysis for differentiation of large adrenal cortical tumours on CT. *Clin Radiol* (2019) 74:818.e1–7. doi: 10.1016/j.crad.2019.06.021

24. Ho LM, Samei E, Mazurowski MA, Zheng Y, Allen BC, Nelson RC, et al. Can texture analysis be used to distinguish benign from malignant adrenal nodules on unenhanced CT, contrast-enhanced CT, or in-phase and opposed-phase MRI? *AJR Am J Roentgenol* (2019) 212:554–61. doi: 10.2214/AJR.18.20097

25. Koyuncu H, Ceylan R, Asoglu S, Cebeci H, Koplay M. An extensive study for binary characterisation of adrenal tumours. *Med Biol Eng Comput* (2019) 57:849–62. doi: 10.1007/s11517-018-1923-z

26. Li X, Guindani M, Ng CS, Hobbs BP. Spatial Bayesian modeling of GLCM with application to malignant lesion characterization. *J Appl Stat* (2018) 46:230–46. doi: 10.1080/02664763.2018.1473348

27. Liu J, Xue K, Li S, Zhang Y, Cheng J. Combined diagnosis of whole-lesion histogram analysis of T1- and T2-weighted imaging for differentiating adrenal adenoma and pheochromocytoma: A support vector machine-based study. *Can Assoc Radiol J = J l'Association Can Des Radiol* (2021) 72:452–9. doi: 10.1177/0846537120911736

28. Moawad AW, Ahmed A, Fuentes DT, Hazle JD, Habra MA, Elsayes KM. Machine learning-based texture analysis for differentiation of radiologically indeterminate small adrenal tumors on adrenal protocol CT scans. *Abdom Radiol (New York)* (2021) 46:4853–63. doi: 10.1007/s00261-021-03136-2

29. Rocha TO, Albuquerque TC, Nather JCJ, Salmon CEG, Tucci SJ, Wang JZ, et al. Histogram analysis of adrenal lesions with a single measurement for 10th percentile: Feasibility and incremental value for diagnosing adenomas. *AJR Am J Roentgenol* (2018) 211:1227–33. doi: 10.2214/AJR.17.19159

30. Romeo V, Maurea S, Cuocolo R, Petretta M, Mainenti PP, Verde F, et al. Characterization of adrenal lesions on unenhanced MRI using texture analysis: A machine-learning approach. *J Magn Reson Imaging* (2018) 48:198–204. doi: 10.1002/jmri.25954

31. Schieda N, Krishna S, McInnes MDF, Moosavi B, Alrashed A, Moreland R, et al. Utility of MRI to differentiate clear cell renal cell carcinoma adrenal metastases from adrenal adenomas. *AJR Am J Roentgenol* (2017) 209:W152–9. doi: 10.2214/AJR.16.17649

32. Shi B, Zhang G-M-Y, Xu M, Jin Z-Y, Sun H. Distinguishing metastases from benign adrenal masses: what can CT texture analysis do? *Acta Radiol* (2019) 60:1553–61. doi: 10.1177/0284185119830292

33. Shoemaker K, Hobbs BP, Bharath K, Ng CS, Baladandayuthapani V. Tree-based methods for characterizing tumor density heterogeneity. *Pac Symp Biocomput* (2018) 23:216–27. doi: 10.1142/9789813235533_0020

34. Stanzione A, Cuocolo R, Verde F, Galatola R, Romeo V, Mainenti PP, et al. Handcrafted MRI radiomics and machine learning: Classification of indeterminate solid adrenal lesions. *Magn Reson Imaging* (2021) 79:52–8. doi: 10.1016/j.mri.2021.03.009

35. Szász P, Kučera P, Čtvrtlík F, Langová K, Hartmann I, Tüdös Z, et al. Diagnostic value of unenhanced CT attenuation and CT histogram analysis in differential diagnosis of adrenal tumors. *Medicina-Lithuania* (2020) 56(11):597. doi: 10.3390/medicina56110597

36. Torresan F, Crimì F, Ceccato F, Zavan F, Barbot M, Lacognata C, et al. Radiomics: a new tool to differentiate adrenocortical adenoma from carcinoma. *BJS Open* (2021) 5(1):zraa061. doi: 10.1093/bjsopen/zraa061

37. Tu W, Verma R, Krishna S, McInnes MDF, Flood TA, Schieda N. Can adrenal adenomas be differentiated from adrenal metastases at single-phase contrast-enhanced CT? *AJR Am J Roentgenol* (2018) 211:1044–50. doi: 10.2214/AJR.17.19276

38. Tu W, Abreu-Gomez J, Udare A, Alrashed A, Schieda N. Utility of T2-weighted MRI to differentiate adrenal metastases from lipid-poor adrenal adenomas. *Radiol Imaging Cancer* (2020) 2:e200011. doi: 10.1148/rycan.2020200011

39. Tüdös Z, Kučera P, Szász P, Hartmann I, Langová K, Škarda J, et al. Influence of slice thickness on result of CT histogram analysis in indeterminate adrenal masses. *Abdom Radiol (New York)* (2019) 44:1461–9. doi: 10.1007/s00261-018-1835-2

40. Umanodan T, Fukukura Y, Kumagae Y, Shindo T, Nakajo M, Takumi K, et al. ADC Histogram analysis for adrenal tumor histogram analysis of apparent diffusion coefficient in differentiating adrenal adenoma from pheochromocytoma. *J Magn Reson Imaging* (2017) 45:1195–203. doi: 10.1002/jmri.25452

41. Wu ZJ, Clark TJ, Hippe DS, Soloff EV, Zhang M, Hsu L, et al. Comparison of histogram-based Gaussian analysis with and without noise correction for the characterization of indeterminate adrenal nodules. *AJR Am J Roentgenol* (2020) 215:896–902. doi: 10.2214/AJR.19.22531

42. Yi X, Guan X, Chen C, Zhang Y, Zhang Z, Li M, et al. Adrenal incidentaloma: machine learning-based quantitative texture analysis of unenhanced CT can effectively differentiate sPHEO from lipid-poor adrenal adenoma. *J Cancer* (2018) 9:3577–82. doi: 10.7150/jca.26356

43. Yi X, Guan X, Zhang Y, Liu L, Long X, Yin H, et al. Radiomics improves efficiency for differentiating subclinical pheochromocytoma from lipid-poor adenoma: a predictive, preventive and personalized medical approach in adrenal incidentalomas. *EPMA J* (2018) 9:421–9. doi: 10.1007/s13167-018-0149-3

44. Yu H, Parakh A, Blake M, McDermott S. Texture analysis as a radiomic marker for differentiating benign from malignant adrenal tumors. *J Comput Assist Tomogr* (2020) 44:766–71. doi: 10.1097/RCT.0000000000001051

45. Zhang G-M-Y, Shi B, Sun H, Jin Z-Y, Xue H-D. Differentiating pheochromocytoma from lipid-poor adrenocortical adenoma by CT texture analysis: feasibility study. *Abdom Radiol (New York)* (2017) 42:2305–13. doi: 10.1007/s00261-017-1118-3

46. Zheng Y, Liu X, Zhong Y, Lv F, Yang H. A preliminary study for distinguish hormone-secreting functional adrenocortical adenoma subtypes using multiparametric CT radiomics-based machine learning model and nomogram. *Front Oncol* (2020) 10:570502. doi: 10.3389/fonc.2020.570502

47. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol* (2016) 61:R150–66. doi: 10.1088/0031-9155/61/13/R150

48. Moon SH, Kim J, Joung J-G, Cha H, Park W-Y, Ahn JS, et al. Correlations between metabolic texture features, genetic heterogeneity, and mutation burden in patients with lung cancer. *Eur J Nucl Med Mol Imaging* (2019) 46:446–54. doi: 10.1007/s00259-018-4138-5

49. Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton JD, Snyder A, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin Radiol* (2017) 72:3–10. doi: 10.1016/j.crad.2016.09.013

50. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging* (2020) 11:91. doi: 10.1186/s13244-020-00887-2

51. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *Eur Radiol* (2020) 30:523–36. doi: 10.1007/s00330-019-06360-z

52. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* (2015) 350:g7594. doi: 10.1136/bmj.g7594

53. Ursprung S, Beer L, Bruining A, Woitek R, Stewart GD, Gallagher FA, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. *Eur Radiol* (2020) 30:3558–66. doi: 10.1007/s00330-020-06666-3

54. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* (2017) 14:169–86. doi: 10.1038/nrclinonc.2016.162

55. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (remark). *Exp Oncol* (2006) 28:99–105. doi: 10.1007/s10549-006-9242-8

56. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* (2020) 295:328–38. doi: 10.1148/radiol.2020191145

57. Jeong E, Park J, Lee J. Diagnostic test accuracy of the 4AT for delirium detection: A systematic review and meta-analysis. *Int J Environ Res Public Health* (2020) 17(20):7515. doi: 10.3390/ijerph17207515

58. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HCW, van der Windt DAWM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* (2002) 2:9. doi: 10.1186/1471-2288-2-9

59. Davnall F, Yip CSP, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging* (2012) 3:573–89. doi: 10.1007/s13244-012-0196-6

60. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* (2018) 2:36. doi: 10.1186/s41747-018-0068-z

61. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* (2015) 349:255–60. doi: 10.1126/science.aaa8415

62. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* (2015) 5:13087. doi: 10.1038/srep13087

63. Mayo-Smith WW, Song JH, Boland GL, Francis IR, Israel GM, Mazzaglia PJ, et al. Management of incidental adrenal masses: A white paper of the ACR incidental findings committee. *J Am Coll Radiol* (2017) 14:1038–44. doi: 10.1016/j.jacr.2017.05.001

64. Crimì F, Quaia E, Cabrelle G, Zanon C, Pepe A, Regazzo D, et al. Diagnostic accuracy of CT texture analysis in adrenal masses: A systematic review. *Int J Mol Sci* (2022) 23(2):637. doi: 10.3390/ijms23020637

65. Stanzione A, Galatola R, Cuocolo R, Romeo V, Verde F, Mainenti PP, et al. Radiomics in cross-sectional adrenal imaging: A systematic review and quality assessment study. *Diagnostics* (2022) 12(3):578. doi: 10.3390/diagnostics12030578

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fonc.2022.961779

OPEN ACCESS

EDITED BY
Xiongbing Zu,
Xiangya Hospital, Central South University, China

REVIEWED BY
Tiezheng Qi,
Xiangya School of Medicine, Central South University, China
Shicheng Guo,
University of Wisconsin-Madison, United States

*CORRESPONDENCE
Ming Liu
csmliu@uestc.edu.cn
Bin Song
songbin@wchscu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 05 June 2022
ACCEPTED 29 August 2022
PUBLISHED 28 September 2022

CITATION
Yang M, He X, Xu L, Liu M, Deng J, Cheng X, Wei Y, Li Q, Wan S, Zhang F, Wu L, Wang X, Song B and Liu M (2022) CT-based transformer model for non-invasively predicting the Fuhrman nuclear grade of clear cell renal cell carcinoma.
Front. Oncol. 12:961779.
doi: 10.3389/fonc.2022.961779

CT-based transformer model for non-invasively predicting the Fuhrman nuclear grade of clear cell renal cell carcinoma

Meiyi Yang[1,2†], Xiaopeng He[3†], Lifeng Xu[5], Minghui Liu[2], Jiali Deng[2], Xuan Cheng[2], Yi Wei[4], Qian Li[4], Shang Wan[4], Feng Zhang[5], Lei Wu[2], Xiaomin Wang[2], Bin Song[4*] and Ming Liu[5*]

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, [3]Department of Radiology, Affiliated Hospital of Southwest Medical University, Luzhou, China, [4]Department of Radiology, West China Hospital, Sichuan University, Chengdu, China, [5]Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China

Background: Clear cell Renal Cell Carcinoma (ccRCC) is the most common malignant tumor in the urinary system and the predominant subtype of malignant renal tumors with high mortality. Biopsy is the main examination to determine ccRCC grade, but it can lead to unavoidable complications and sampling bias. Therefore, non-invasive technology (e.g., CT examination) for ccRCC grading is attracting more and more attention. However, noise labels on CT images containing multiple grades but only one label make prediction difficult. However, noise labels exist in CT images, which contain multiple grades but only one label, making prediction difficult.

Aim: We proposed a Transformer-based deep learning algorithm with CT images to improve the diagnostic accuracy of grading prediction and to improve the diagnostic accuracy of ccRCC grading.

Methods: We integrate different training models to improve robustness and predict Fuhrman nuclear grade. Then, we conducted experiments on a collected ccRCC dataset containing 759 patients and used average classification accuracy, sensitivity, specificity, and AreaUnderCurve as indicators to evaluate the quality of research. In the comparative experiments, we further performed various current deep learning algorithms to show the advantages of the proposed method. We collected patients with pathologically proven ccRCC diagnosed from April 2010 to December 2018 as the training and internal test dataset, containing 759 patients. We propose a transformer-based network architecture that efficiently employs convolutional neural networks (CNNs) and self-attention mechanisms to extract a persuasive feature automatically. And then, a nonlinear classifier is applied to classify. We integrate different training models to improve the accuracy and robustness of

the model. The average classification accuracy, sensitivity, specificity, and area under curve are used as indicators to evaluate the quality of a model.

Results: The mean accuracy, sensitivity, specificity, and Area Under Curve achieved by CNN were 82.3%, 89.4%, 83.2%, and 85.7%, respectively. In contrast, the proposed Transformer-based model obtains a mean accuracy of 87.1% with a sensitivity of 91.3%, a specificity of 85.3%, and an Area Under Curve (AUC) of 90.3%. The integrated model acquires a better performance (86.5% ACC and an AUC of 91.2%).

Conclusion: A transformer-based network performs better than traditional deep learning algorithms in terms of the accuracy of ccRCC prediction. Meanwhile, the transformer has a certain advantage in dealing with noise labels existing in CT images of ccRCC. This method is promising to be applied to other medical tasks (e.g., the grade of neurogliomas and meningiomas).

KEYWORDS

tumor grading, ensemble learning, clear cell renal cell carcinoma, transformer network, deep learning

1 Introduction

Renal cell carcinoma (RCC) is the most common kidney tumor (1, 2). Clear cell RCC (ccRCC) is the predominant hypotype of RCC, accounting for about 75-80 (3). With the background of population aging, kidney cancer, especially RCC, keeps rising. When ccRCC progresses to intermediate and advanced stages, lymph node metastasis or distant organ metastasis probably occurs, which leads to dangerous clinical symptoms and a poor prognosis (4, 5). At present, one of the most important pieces of clinical evidence for judging the malignancy degree of ccRCC is given by the Fuhrman grading system, which defines four pathological grades based on nuclear size, shape, staining, and nucleoli (5). Generally, tumors with low invasiveness are classified as grades I–II, and those with high invasiveness are classified as grades III–IV (6).

The preoperative biopsy is the gold standard for evaluating the grade of ccRCC. However, patients undergoing biopsy are at risk of complications, e.g., hematuria (with more than 80% incidence), perirenal hematoma (with about 60%–90% incidence), and infection (7). The procedure of preoperative biopsy is complex and invasive. Besides, biopsy cannot reflect the Fuhrman grade of the whole tumor (8) because of the high spatial and temporal heterogeneity. Thus, preoperative evaluation of ccRCC using a noninvasive procedure for clinical diagnosis is urgently needed. CT examination is the most commonly used non-invasive technique for preoperative diagnosis and follow-up and plays an essential role in diagnosing and treating renal carcinoma, e.g., detection, localization, characterization, and grading of lesions. In some studies, it has been used to evaluate the preoperative ccRCC classification, such as (9–12).

Preoperative noninvasive prediction of ccRCC is conducive to delivering an individualized treatment. Previous studies (12, 13), using radiation characteristics based on multiphase CT, investigated the predictive performance of different machine learning models for discriminating ccRCC. Beyond that (14–17), have shown that convolutional neural networks based on single or multiphase CT images are beneficial for evaluating ccRCC grading. However, the biggest challenge of CT image grading is the existence of noise labels in the image (8). What is a "noise label?" One CT image may contain multiple grades but may have only one label. For example, the grade III–IV grade CT image contains grade I–II tumor areas because the label was obtained from the biopsy of the most severe tumor area of the whole kidney. When encountering noise labels, convolutions uniformly process all tumor regions regardless of their importance, which leads to the inefficiency of classification. CNN makes decisions based on the convolution kernel, which only focuses on the local pixel subset, resulting in the network tending to learn the local mode rather than the global context.

The transformer network, a branch of deep learning, is considered a promising technology for analyzing medical images because it can capture global representations and establish long-distance relationships within the image (18, 19). Therefore, the transformer is suitable for handling CT images

with noise labels. We propose a transformer classifier, TransResNet, to predict high-grade ccRCC using CT images. To the best of our knowledge, there have been no investigations about discriminating the low and high nuclear grade ccRCC by combining transformer network and radiological features.

As a result, this research aims to investigate an efficient transformer classifier for predicting the Fuhrman grading of ccRCC based on three-phased CT images.

2 Materials and methods

2.1 Data preparation

Patient cohort: This diagnostic and observational study was approved by the institutional review board (West China Hospital, Sichuan University), and written informed consent was obtained from all patients. Consecutive patients were collected from April 2010 to December 2018 in one hospital. We cleaned the original data according to the following rules: (1) there is no obvious noise in the image of the patient; (2) the patient image has no apparent artifacts. Specifically, all of the pathologically proven ccRCC grades are reconfirmed by experienced radiologists. This work uses the Fuhrman nuclear grading system as the standard grading system. Finally, 759 patients were included in this work.

CT images: All patients underwent a multi-slice CT scan with three phases, including unenhanced, arterial, and portal venous phases, using the following systems: LightSpeed VCT (GE Healthcare), Sensation 64 CT (Siemens), or Sensation 16 CT (Siemens). The PCP, CMP, and NP of the MDCT (multidetector CT) examination were acquired for each patient with the following protocol. By using a high-pressure injector at a rate of 3.5 ml/s, 70–100 ml of contrast agents were injected into the antecubital vein. The CMP is the corticomedullary phase contrast-enhanced scan starting 30 s after the contrast agent injection. The NP is the nephrographic phase contrast-enhanced scan starting 90 s after the contrast materials injection. Spiral scanning and thin-slice reconstruction were used for all three phases. The CT scanning parameters of the phases were as follows: the tube voltage was 120 KV; the reconstruction thickness was 1 mm to 5 mm, and the matrix was 512 × 512. All CT scans of patients are converted to color-scale images and reviewed by experienced radiologists in abdominal imaging. The ccRCC images collected from the CMP phase are used in experiments.

CT Processing: Medical CT slices are diverse and complex. If CT slices are selected to be the input of the classification model, the results are imperfect and require further optimization. To make a more accurate diagnosis, we preprocess the original CT image by detecting the organ or lesion area from medical scanning. We utilize a tumor detection network to quickly and efficiently obtain the rectangular region of interest of the tumor in each phase image. As shown in Figure 1, to reduce the complexity of direct tumor segmentation, the detection frame is divided into two stages: (1) renal organ detection: this detection module is composed of VGG16 (20) without the classification layers (pre-trained on ImageNet (21)), aiming to find the rectangular region of kidneys to mitigate the influence of background of CT scans and reduce the search space of tumors; (2) tumor detection: aims to regress the rectangular region of the tumor accurately and predict the possibility of the tumor.

Because medical data are scarce and difficult to label, large deep learning models rely on data augmentation to improve performance. To study the impact of data enhancement, we will consider several common enhancements here. There are two types of data augmentation in the vision computer domain. One is appearance transformation, such as sharpness, brightness, contrast, saturation, gray processing, Gaussian blur, and elasticity (22), another involves spatial geometric transformation, such as horizontal flipping, rotation, cropping, and resizing (23). Every enhancement strategy can transform data stochastically with internal parameters (e.g., rotation degree, noise level). Our model adopted strategies including random clipping, Affine, Gaussian blur, and Gaussian noise. Figure 2 visualizes the augmentations that we study in this work.

2.2 ccRCC classification network

To make the features extracted from the ccRCC dataset correspond to the label as much as possible, we introduced a transformer module into convolutional networks to improve performance. We illustrate the overall diagram in Figure 3A. Firstly, the input image is processed with several convolution blocks, and then the processed feature map is provided to the transformer block. At the end of the network, the class tokens are applied for the prediction of ccRCC. Our insights into taking advantage of convolutions and transformers are: (1) in the early network, using convolution to learn densely distributed and low-level features requires less computational cost than the transformer; and (2) in the later network, applying the transformer to learn higher-level semantic concepts and long-range dependencies in the image.

The network framework, termed TransResNet, is composed of convolution blocks, transformer blocks, and classifier (one FC layer). The convolution block adopts a pyramid structure, in which the resolution of feature mapping decreases with the increase in network depth while the number of channels increases. It consists of four phases: the first phase is a 3 × 3 convolution with stride 1 and padding 1, which is used to extract initial local features. In the last three stages, we refer to the first three layers of ResNet, in which the output channels of each

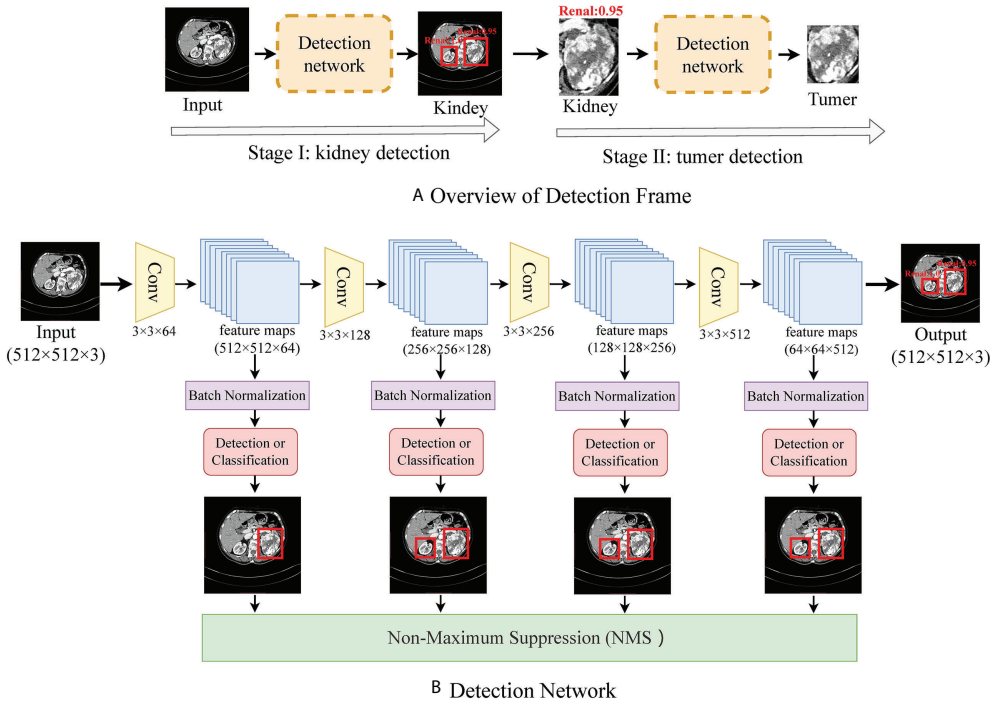


FIGURE 1
(A) shows the flow frame of data processing, in which stage one is the kidney detection network and stage two is a tumor detection network. **(B)** shows the details of the detection network.

layer are 16, 32, and 64, respectively. The Transformer block contains 12 repeated transformers. As shown in Figure 3B, each transformer consists of a multi-head self-attention (MSA) module and a multi-layer perceptron (MLP) block.

In the Transformer, we consider the input feature map $X \in \mathcal{R}^{c \times h \times w}$ with c channels and the feature shape of $h \times w$. Self-attention estimates the relationship between one part of a feature map and other parts (e.g., which tumor masses are likely to come together into a complete tumor mass with maximum grade). Therefore, the feature map is divided into a sequence $X = \{x_1, x_2, \ldots, x_n\}$ $(x \in R^{c \times \frac{h}{n} \times \frac{w}{n}})$ with n patches. The goal of self-attention is to capture the interaction among all n patches by encoding each patch in terms of global contextual information. The output of the MSA layer is computed with the feature map X using the following equation:

$$Z = MSA\left(LN\left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V + X_{class} + X_{position}\right)\right) + X \quad (1)$$

where Queries W^Q , Keys W^K, and Value W^v are learnable weights to automatically learn the importance of each patch. The

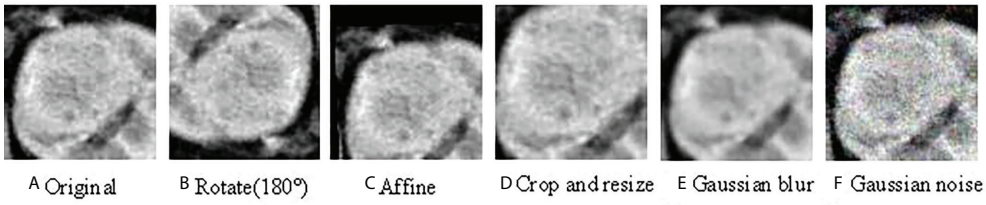


FIGURE 2
Illustrations of the results using various data augmentation **(A)** shows the original CT image; **(B)** is the CT image after 180° rotation; **(C)** is the CT image after Affine transformation; **(D)** shows the CT image after crop and resize transformation; **(E)** is the CT image with Gaussian blur; **(F)** is the CT image with Gaussian noise.

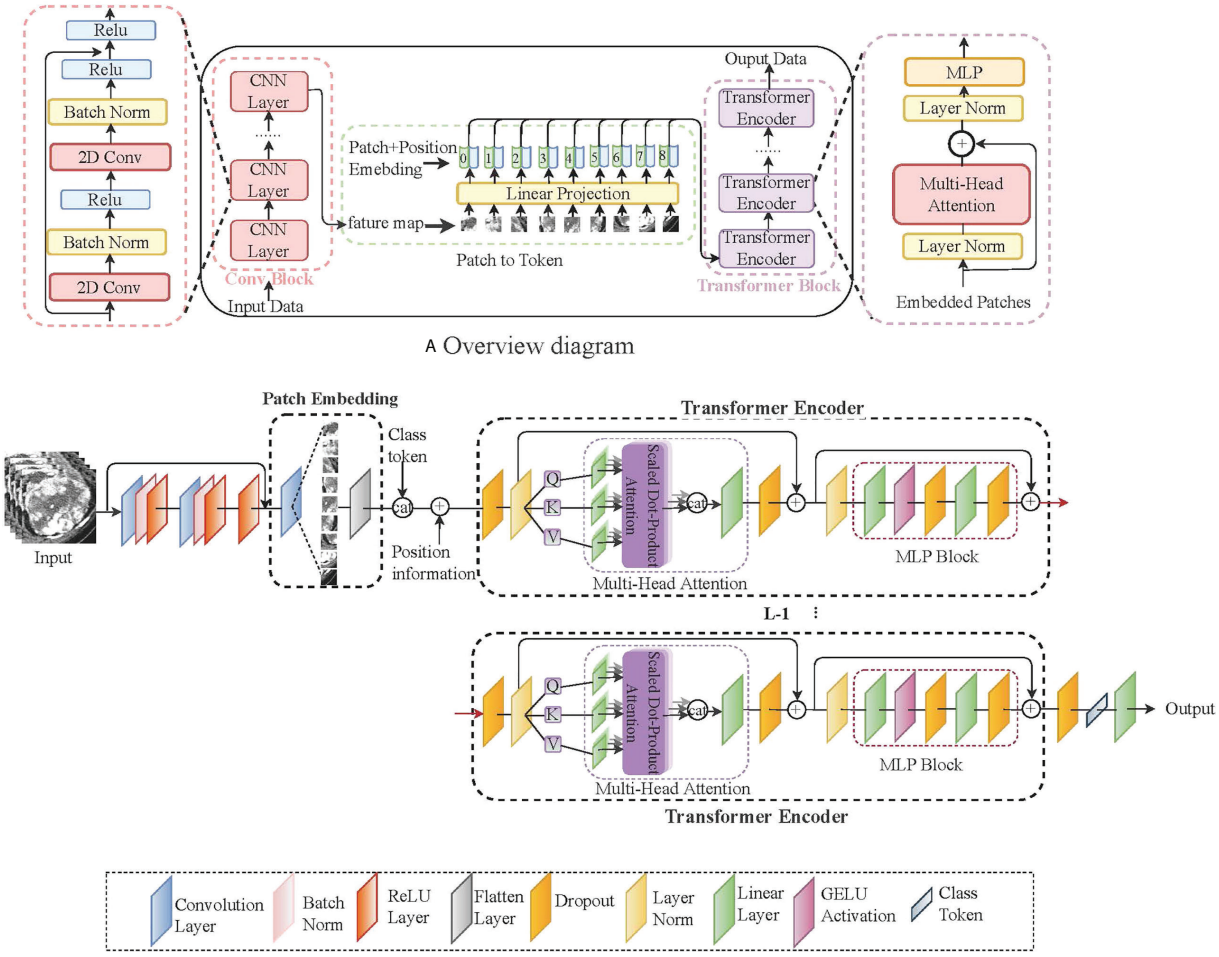


FIGURE 3
A simple network framework for TransResNet. **(A)** represents the overall architecture, mainly including the CNN structure and the transformed structure. **(B)** shows the details of TransResNet. The network framework contains 12 transformer blocks (i.e., L = 12).

input sequence X is projected onto these weight matrices to get $Q=XW^Q$, $K=XW^K$ and $V=XW^V$. The d_k is the dimension of key vector k, providing an appropriate normalization to make the gradient more stable. LN represents the linear normalization, X_{class} and $X_{position}$ are randomized parameters representing classification and position information, respectively. To capture the structure of object image, position and classification information are merged into the self-attention feature map for training.

After the self-attended feature map Z passes through the linear layer, we put it into a MLP including two linear transformations and a GELU activation to perform the same operation on the vector of each position. The output can be obtained:

$$Y = MLP(LN(Z)) + Z \tag{2}$$

Referring to ViT (24), we take the classification information X_{class} instead of extracted image features as the input of the classifier FC to classify directly. We train our network using the cross-entropy (CE) loss between prediction and ground truth, which can be written as:

$$\mathcal{L}_{CE} = -\sum \left(y \log \hat{y} + (1-y) \log (1-\hat{y})\right) \tag{3}$$

where the y is the real label and $\hat{y}$ is the predicted probability. The loss function represents the difference between the real label and the predicted probability.

We constructed TransResNet with three residual blocks and 12 transformer blocks. The major hyper-parameters are as follows: the optimizer is Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5e−4. The batch size is 100 per worker. For the

epochs, the learning rate is scaled linearly from 0.01 to 0.00001, and then it is divided by 10 at epochs 50, 100, and 150. The proposed model is implemented using Pytorch 1.0.1. We ran the experiments on an Ubuntu 16.04.3 server with four NVIDIA GeForce GTX 1080 cards. Specific details of the code can be seen at: https://github.com/yangmeiyi/ccRCC_project.git.

2.3 Model ensembles

To a certain degree, ensemble learning can improve prediction accuracy. To achieve a strong classifier, we train multiple different classifiers. We selected the following five networks as sub-networks of the integration model:

- TransResNet: In medical, ResNet (25) is the preferred model because it is simple and efficient. The most common model, ResNet50, is ineffective on our dataset due to serious over-fitting. Finally, we decided to pair Transformer and resnet34.
- TransDenseNet: DenseNet (26) found that some layers are randomly lost at each step in the training process, which can significantly improve the generalization performance of ResNet. Similarly, we chose to combine Transformer and DenseNet-121 with the smallest model parameters as our sub-network.
- TransInception: Compared with the structure of ResNet, Inception (27) not only increases the width of the network but also embeds features of different scales. Similarly, Inception-V3 was selected by us.
- TransSENet: Different from the previously mentioned networks, improving the performance through spatial latitude, SENet (28) establishes the interdependence between feature channels.
- TransRegNet: RegNet (29) proposes a new design paradigm that estimates the overall network design space (depth and width) to obtain the best design. Similarly, we chose RegNetY-200MF with the minor model parameters.

We combine them through the average method. These classifiers and TransResNet are constructed in the same way. The difference lies in the construction of convolution blocks, such as TransDenseNet uses a dense convolutional network and TransInception adopts an inception network.

Suppose we have got N trained different classifiers. Facing a test sample x ($x \in$ test dataset), the prediction results of the integrated model are a N-dimensional vector $\{S_1,S_2,\ldots,S_n\}$. The final score using averaging is formulated as follows:

$$Final\_score = \frac{\sum_{i=1}^{N} S_i}{N} \quad (4)$$

3 Results

3.1 Clinical characteristics

This study included 759 patients, which comprised 477 low-grade [grade I (n = 25, 5.24%) and II (n = 452, 94.76%)] ccRCC patients and 282 high-grade [grade III (n = 232, 82.27%) and IV (n = 50, 17.73%)] ccRCC patients. Male and female patients are equally represented, accounting for 64% and 46%, respectively. The patient characteristics of the training and testing cohorts are shown in Table 1.

3.2 Performance of proposed algorithm

To validate this model, we choose Accuracy (ACC), Area Under Curve (AUC), Sensitivity (SE), and Specificity (SP) as evaluation criteria.

Performance with data enhancement: To comprehend the role of enhancement strategy in detail, the individual and different combinations of them are discussed. With inconsistent images of the ccRCC, cropping as the basic data processing is applied. Figure 4 shows the evaluation results under single and combined transformations. Even though Affine is a very effective enhancement for the model, any single transformation is insufficient for representation learning. One augmentation composition stands out: random Affine and random Gaussian noise. The best combination is more than two enhancements, such as random Rotate, random Affine, and random Crop, making the model obtain 87% ACC and 91% AUC.

Performance on different architectures. This section shows the effect of the transformer on ResNet (25), DenseNet (26), Inception (27), SENet (28), and RegNet (29). Each model has its own unique advantages. We present the ACC, AUC, SE, and SP of TransResNet, TransDenseNet, TransInception, TransRegNet, and TransSENet on the ccRCC dataset in Table 2. Each architecture was performed under the enhancement method of

TABLE 1 The demographic and clinical statistics of patients with ccRCC.

Attribute	Training cohort	Testing cohort
Age (years)	56 ± 12 (589)	54 ± 11 (170)
Male	374	112
Female	215	58
Grade I	21	4
Grade II	371	81
Grade III	165	67
Grade IV	32	18

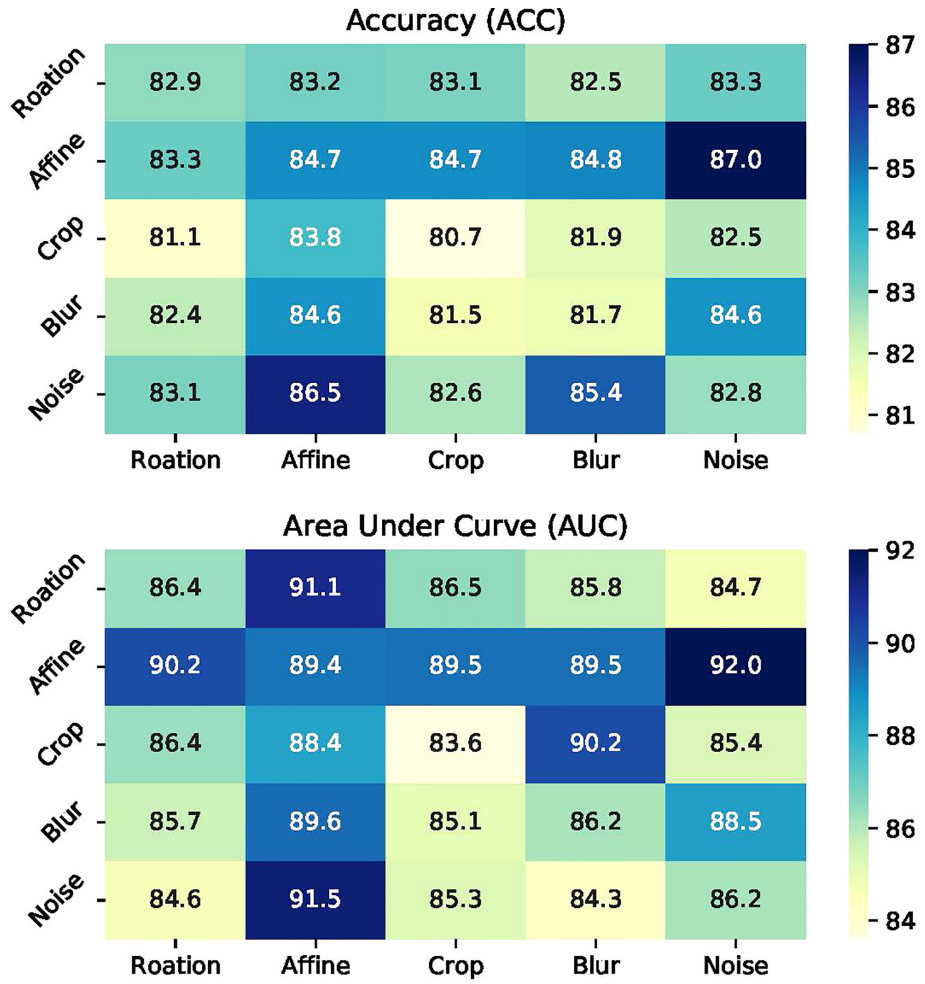


FIGURE 4
Illustrations of data augmentation operators.

Crop, random Horizontal Flip, random Rotation, and random Affine.

Table 2 shows the experimental results of different network architectures. The Receiver Operating Characteristic (ROC) curve is a metric that can provide a pure index of accuracy, which has been accepted by every researcher and applied in medical studies. We show the ROC curve in Figure 5.

Ensemble results: Ensemble learning is a popular way to improve robustness and accuracy by training a group of heterogeneous models. These heterogeneous models are combined by different strategies such as voting, averaging, stacking, and blending. This study analyzes various ensemble strategies, and the results are shown in Table 3. The integrated model consists of a single model: TransResNet, TransDenseNet, TransInception, TransRegNet, and TransSENet.

3.3 Comparison with SOTA methods

Does our proposed solution lead to better performance than pure transformers? To answer this question, we compared pure transformer networks ViT (24) and CaiT (30) with the smallest parameters and a hybrid network Conformer (31) trained in parallel by convolution and transformer in Table 4. In the training process, ViT-Small and CaiT-Small are suitable for our dataset, but the results are still unsatisfactory. In particular, the hybrid structure Conformer with huge model parameters is difficult to train on small dataset ccRCC, resulting in the worst performance, although its classification results on ImageNet are better than ViT and CaiT.

Does the transformer improve performance compared to CNN? For CNN, we have done a series of experiments with ccRCC for comparison. The experiments included Data

TABLE 2 The impact of transformer on different network structures. .

Model	ACC	AUC	SE	SP
ResNet-34	82.3 ± 2.5	85.7 ± 2.3	89.4 ± 1.7	83.2 ± 1.2
TransResNet	87.1 ± 2.3	90.3 ± 2.5	91.3 ± 1.4	85.3 ± 1.5
DenseNet-121	81.5 ± 2.3	85.5 ± 2.4	80.0 ± 0.5	84.3 ± 1.2
TransDenseNet	83.9 ± 2.1	90.5 ± 2.2	80.6 ± 0.6	86.8 ± 1.0
Inception-V3	80.0 ± 2.0	83.7 ± 2.0	76.5 ± 1.2	78.8 ± 1.3
TransInception	84.3 ± 2.0	89.4 ± 2.0	83.4 ± 1.3	85.8 ± 1.6
SENet	81.8 ± 2.5	84.2 ± 2.8	76.8 ± 1.5	85.1 ± 1.4
TransSENet	85.1 ± 2.3	89.2 ± 2.5	89.1 ± 1.3	82.1 ± 1.1
RegNet	81.9 ± 2.1	84.3 ± 2.5	82.5 ± 1.5	80.0 ± 1.0
TransRegNet	82.3 ± 2.0	87.7 ± 2.5	84.7 ± 1.0	80.0 ± 0.5

We report the average accuracy and standard deviation of five time runs.
ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively.

Enhancement, CNN architecture, Transfer Learning, and the Regularization Method. The experimental results are shown in Figure 6. In the experiment of data enhancement, the performance of the CNN model has definite improvements after random Affine. However, other data enhancement methods are not effective for the dataset. We found that the CNN architectures have little effect on the Fuhrman Grade of ccRCC, although the advantages of each structure are different. From various experimental results, the model based on CNN is difficult to break through the bottleneck of AUC (e.g., 85 AUC). Our method, TransResNet, easily surpasses CNN algorithms using various tricks in ACC and AUC.

3.4 Comparison with transfer learning

Transfer learning is widely used in medical image processing, such as type 2 diabetes (32), 3D genome folding (33), and papillary thyroid carcinoma (34), etc. Many factors make it impossible to establish large-scale datasets such as ImageNet in the medical field, so the limited data restricts the performance of deep neural networks. Some studies reckon that the pre-trained model obtained on ImageNet can be fine-tuned on the medical dataset to acquire high performance. The experiment by Shin et al. proved that although there are differences between natural images and CT images, CNN fully trained on large-scale well-annotated datasets may still be

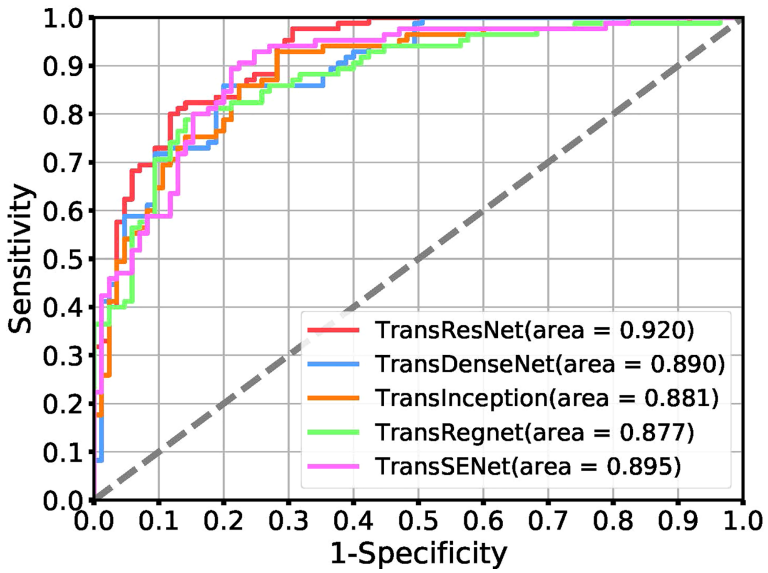


FIGURE 5
Receiver operating characteristic (ROC) curves for the task of tumor classification using a positive ratio feature.

TABLE 3 Ensemble results under different strategies.

Model	ACC	AUC	SE	SP
TransResNet	85.8	90.3	91.3	85.3
TransDenseNet	83.9	90.5	80.6	86.8
TransInception	84.3	89.4	83.4	85.8
TransRegNet	82.3	87.7	84.7	80.0
TransSENet	85.1	89.2	89.1	82.1
Ensemble(voting)	85.8	90.5	91.4	89.8
Ensemble(averaging)	**86.5**	**91.2**	**92.1**	**89.5**
Ensemble(stacking)	86.1	90.8	89.1	91.1
Ensemble(blending)	86.1	90.0	90.0	85.5

ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. The averaging strategy of ensemble shows the best results (the seventh column, highlighted in bold).

transmitted to make the medical image recognition task more effective (35). Therefore, we compared the performance of our method and transfer learning on ccRCC. We transferred the ResNet, DenseNet, and ViT trained with ImageNet to ccRCC for fine-tuning, respectively. In transfer learning, to make the network training more thorough, we follow the method of (35): all CNN layers are fine-tuned with a learning rate 10 times lower than the default learning rate, except for the last layer. The last fully connected layer is randomly initialized and newly trained to adapt to the new object categories in the ccRCC application. Its learning rate remains at the original 0.01. The results are shown in Table 5.

In the experiment, we found that although transfer learning can speed up convergence early in training, it does not improve final accuracy. As shown in Table 5, the accuracy of our proposed method can quickly catch up with the best of transfer learning and stay higher.

3.5 Analysis of noise labels

Previous articles (15, 17) studied the impact of CNN-based deep learning on ccRCC. These studies prove that the features from CT images extracted by CNN can be effectively used for ccRCC grading and obtaining the SOTA performance. Due to the inconsistency between extracted features and labels, some data are always misclassified by the CNN-based algorithm. For example, 21 CT images in our test set could not be classified correctly, including three low-grade images and 18 high-grade images. This phenomenon is consistent with the fact that high-grade images contain more noise information. In other words, the features extracted from tumor images scanned by high-grade ccRCC CT scans are not always related to Fuhrman nucleus grades III and IV, and some of the features are related to Foreman nucleus grades I and II. Figure 7 shows the case of misclassification data, in which Figure 7A is a positive sample, but it is classified as a negative example. Figure 7B is a negative example sample of classification error. Figure 7C is a class activation map of a positive example sample of classification error. Figure 7D is a class activation map of a positive example sample of classification error. We found that: (1) similar CT images with different categories are difficult to distinguish (e.g., Figures 7A, B), and (2) some CT images are not activated correctly (e.g., Figures 7C, D).

The Transformer offers many advantages over traditional CNN algorithms for ccRCC grading. For example, Transformer can capture the representation of global images and establish long-distance relationships in CT images, reflecting the microscopic heterogeneity changes of tumors more comprehensively and providing a more accurate diagnosis than all CNN algorithms. In addition, the complementary strengths between CNN and Transformer can be adapted to the requirements of specific clinical environments. After the optimized classifier TransRsenet, 14 CT images in our test set could not be correctly classified, including two low-grade images and 12 high-grade CT images. Figure 8 shows the class activation map of these error data to show the shift of attention of the model during training. Compared with the convolutional network, the model with the proposed method makes objects clearer and more accurate than the original ones.

TABLE 4 Comparison with the state-of-the-art transformer.

Model	ACC	AUC	SE	SP
TransResNet	**85.8**	**90.3**	**91.3**	**85.7**
ViT-Small	78.6	83.5	88.5	81.3
CaiT-Small	79.4	83.2	89.1	82.0
Conformer	76.4	79.3	83.5	65.7

ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. Our model TransResNet obtains the best results (the first column, highlighted in bold).

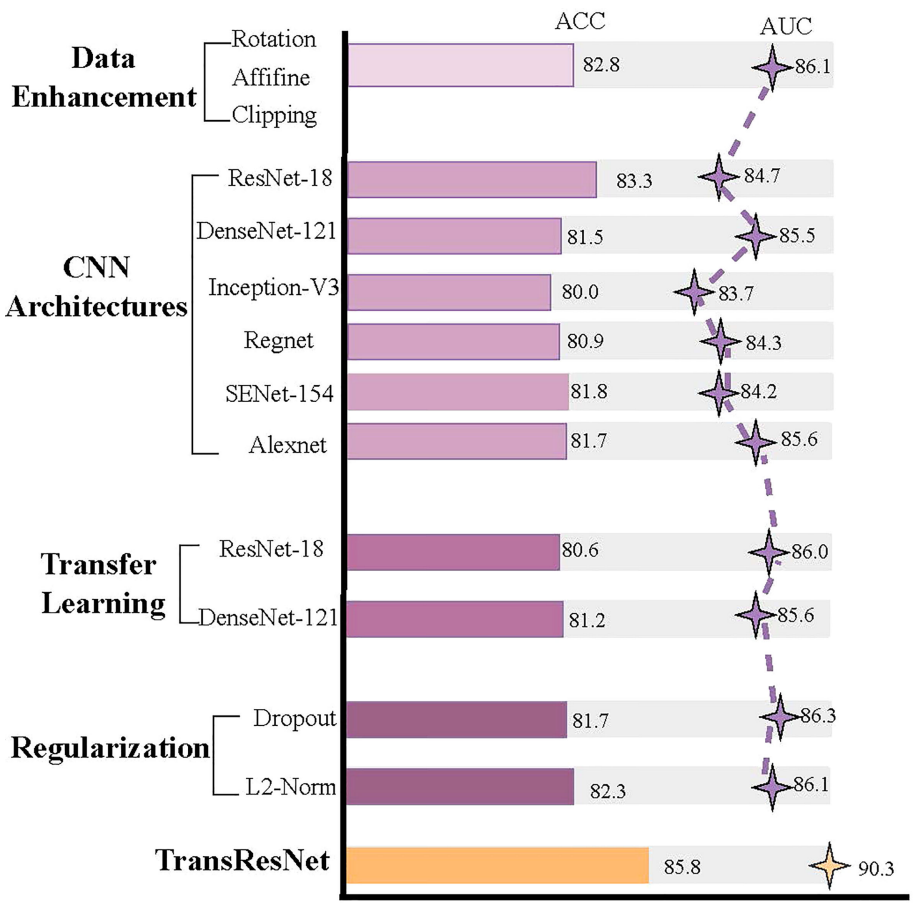


FIGURE 6
A series of experimental results about CNN.

In addition to the comparison of convolution algorithms, we also compared the network architecture of SOTA, such as the transformer-based ViT, CaiT, and Conformer, obtaining the best grading results under ACC, SE, SP, and AUC. In conclusion, we demonstrated the effectiveness of the Transformer module over ccRCC grading. Furthermore, the combination of CNN and Transformer mitigates the noise label problem in ccrRCC.

3.6 External validation

We evaluate our model based on one external validation dataset, the Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) (36, 37). The TCGA-KIRC focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome

TABLE 5 The comparison results of our method and transfer learning.

Transfer Model	ACC	AUC	SE	SP
ResNet18	80.6	86.0	83.3	78.8
ResNet34	80.2	85.6	86.3	78.2
DenseNet121	81.2	85.6	78.9	89.4
TransResNet	**85.8**	**90.3**	**91.3**	**85.3**
ViT-tiny	74.7	75.6	64.7	84.7

The comparison models are ResNet, DenseNet, and ViT-tiny parameterized by pre-training using the ImageNet dataset. Our model is trained from random initialization.
ACC ,AUC ,SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. Our model TransResNet obtains the best results (the seventh column, highlighted in bold).

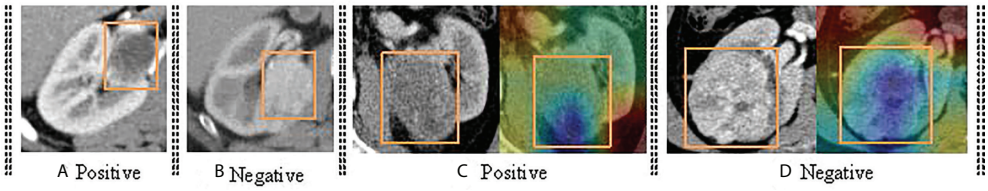


FIGURE 7
The demonstration of error classification. It mainly includes four categories. **(A)** shows the the CT image of positive samples; **(B)** shows the the CT image of negative samples; **(C)** shows the CT image of positive samples and corresponding class activation map; **(D)** shows the CT image of negative samples and corresponding class activation map.

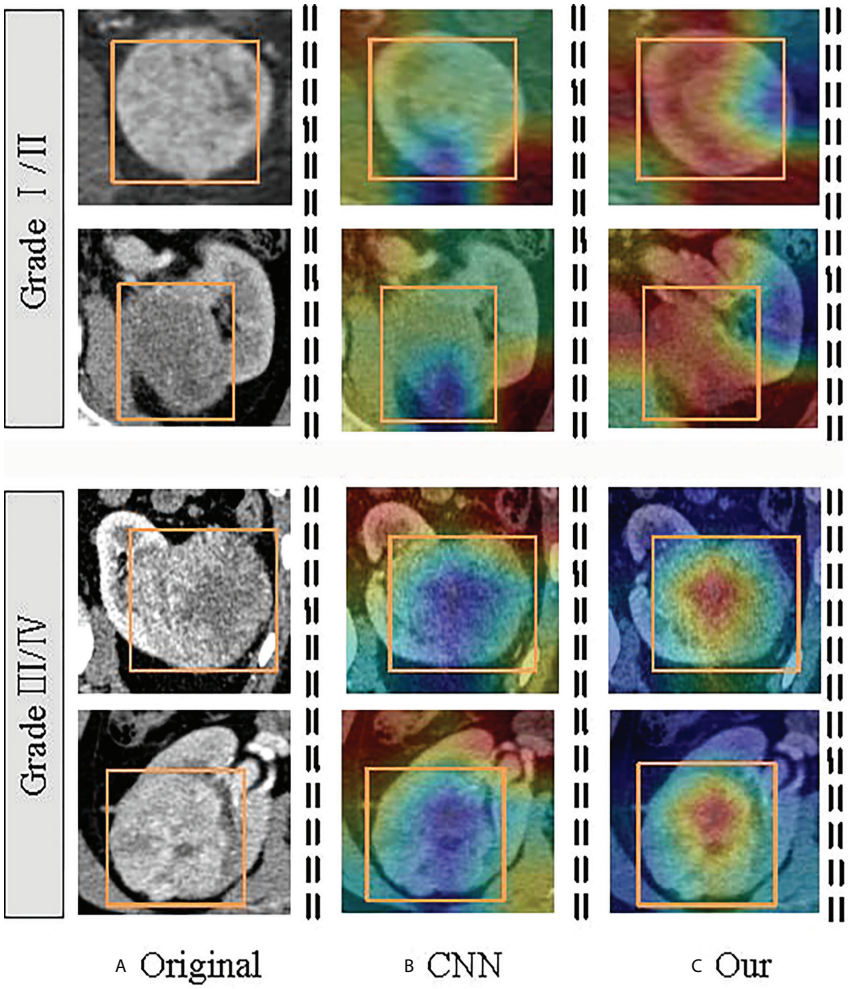


FIGURE 8
Visualization of the class activation map generated by the last transformer layer on images from the ccRCC. The yellow box indicates the lesion area. The color red denotes higher attention values, and the color blue denotes lower. **(A)** is the original image; **(B)** shows the class activation map of the CNN model; and **(C)** is the class activation map of the TransResNet model.

Atlas (TCGA). Clinical, genetic, and pathological data resides in the Genomic Data Commons (GDC) Data Portal, while the radiological data is stored in The Cancer Imaging Archive (TCIA). Here we just use the CT radiological data. We selected 20 patients (e.g., 10 low-grade samples and 10 high-grade samples) from 227 patients as our external validation data. The results of our model TransResNet on the data are: 81.9% ACC, 85.4% AUC, 76.6% SE, and 87.2% SP.

4 Discussion

This retrospective study comprehensively investigates the pros and cons of different deep learning models based on two-phase CT images for discriminating low- and high-grade ccRCC. According to data scarcity and noise labels, a transformer-based deep learning model, TransResNet, is proposed. After a thorough comparison, the proposed discriminative model, TransResNet, achieved satisfactory performance. In addition, we find that this model can effectively alleviate the impact of noise labels.

At present, the analysis of radiomics is a common paradigm to predict the ccRCC. Previous studies analyzed the texture of non-enhanced CT through machine learning to obtain universal features for higher accuracy. For example, Coy et al. have shown that high-grade ccRCC lesions are significantly larger and have more calcifications, necrosis, aggregation system infiltration, and unclear tumor margins than low-grade ccRCC lesions (38). In addition, several studies (39, 40) have proved that high-grade tumors tend to be larger than low-grade. However, our experiment found high-grade lesions are not significantly different from low-grade CT images. This is consistent with the findings of (12) and (13). The performance of image-based quantitative indicators varies on different datasets, and their adequacy needs further verification. In this study, we use deep learning to analyze the CT images, which can automatically discover pixel-level features, supporting a more powerful model.

In ccRCC, some high-grade lesions have low-grade features due to unavoidable sampling bias. It was also interesting to find that convolution-based networks focus on local pixel subsets, which results in a tendency to learn local patterns so that the relationships within images are ignored. As a result, in the complex samples, the convolution kernel cannot effectively make decisions, which leads to unsatisfactory grading accuracy. CNNs uniformly process CT images regardless of their importance. In contrast, the transformer uses a self-attention mechanism instead of a CNN, which can establish long-range dependencies in images. The Transformer has been widely used for medical tasks with high accuracy, e.g., prostatic segmentation (41), delineating the epicardium and endocardium (42), multi-modal medical image classification (18), etc. The Transformer can not only be used for segmentation but also carries an advantage in processing images with noisy labels due to its way of capturing image features. However, transformer networks require a large-scale dataset for training because the transformer lacks inductive bias (24). Thus, scarce medical data leads to severe over-fitting, which may further reduce the reliability of the classifier. Therefore, we take advantage of the CNN and a transformer to extract the features of CT images. The introduced transformer module can help CNN obtain the relationship between the pixel blocks inside the image. We find that this hybrid classifier is superior to a single classifier, with ACC increasing from 83% to 87% and AUC increasing from 85% to 90%.

Correctly grading complex samples is one of the significant indicators for evaluating the strength and stability of a deep learning model, which reflects the ability to extract general features. Our method shows some advantages when faced with noise samples and difficultly graded examples (see *Analysis of noise labels* section). As is known to all, different models have respective preferences for data, and the features learned from a single model are not necessarily reliable. This paper trains a set of heterogeneous models to improve robustness and accuracy by integrating their extracted features, with ACC increasing from 85.8% to 86.5% and AUC increasing from 90% to 91.2%. This study shows that our method can effectively reduce the difficulty of grading caused by noise labels. This method can be widely applied to other medical grading tasks, e.g., the grade of neurogliomas and meningiomas.

Our study has several limitations. First, Fan et al. (12) found that the classifier based on a three-phase CT image is better than that based on a single-phase CT image. Theoretically, non-enhanced CT images can provide additional information for diagnosis. However, this study only used the data from the arterial and portal vein phases due to fewer non-enhanced CT images and inapparent texture information. Additionally, we did not collect additional validation sets because the original DICOM data of the ccRCC are difficult to obtain. However, the superiority of the transformer in the feature extraction can be reflected in the training and testing.

5 Conclusion

This paper studies the application of transformer architecture to ccRCC classification. We first collected a high-quality ccRCC CT scan dataset containing more than 759 patients with pathological proven. Then, a hybrid structure, termed TranResNet, is proposed, which compromises the merits of CNN and Transformer. Unlike other transformer models, TranResNet does not require pre-training on large-scale datasets. Finally, we conducted extensive experiments on ccRCC datasets to verify our method. TransResNet achieves good performance over ConvNets and other related transformer architectures, demonstrating promising results in ccRCC classification. We hope it will help future research on this subject, and it can cooperate with radiologists to classify the ccRCC in an actual clinical situation.

Data availability statement

The original contributions presented in the study are included in the article material, further inquiries can be directed to the corresponding author.

Ethics statement

This diagnostic and observational study was approved by the institutional review board (West China Hospital, Sichuan University). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

MY designed the study and all laboratory studies, performed experiments, analyzed the data, and wrote the manuscript. ML, JD, and XC performed laboratory experiments. XH, YW, QL, and SW collected and analyzed the ccRCC data. LX and FZ provided help in data and statistical analysis. LW and XW participated in data analysis and interpretation. ML and BS oversaw the design of the study and all laboratory studies, data analysis, and interpretation, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2021YGLH213, No. ZYGX2022YGRH016), the Municipal Government of Quzhou (Grant 2021D007, Grant 2021D008, Grant 2021D015, Grant 2021D018), as well as the Zhejiang Provincial Natural Science Foundation of China under Grant No. LGF22G010009.

Conflict of interest

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.

Publisher's note

References

1. Capitanio U, Montorsi F. Renal cancer. *Lancet* (2016) 387(10021):894–906. doi: 10.1016/S0140-6736(15)00046-X

2. Incorvaia L, Procopio G, Porta C. Renal Cancer. In: A Russo, M Peeters, L Incorvaia and C Rolfo (eds) *Practical Medical Oncology Textbook*. UNIPA Springer Series., Cham: Springer. (2021) pp.755–74. doi: 10.1007/978-3-030-56051-5_45

3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA: Cancer J Clin* (2021) 71:7–33. doi: 10.3322/caac.21654

4. Fuhrman SA, Lasky LC, Limas C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol* (1982) 6:655–63. doi: 10.1097/00000478-198210000-00007

5. Delahunt B, Eble JN, Egevad L, Samaratunga H. Grading of renal cell carcinoma. *Histopathology* (2019) 74:4–17. doi: 10.1111/his.13735

6. Feng Z, Shen Q, Li Y, Hu Z. Ct texture analysis: a potential tool for predicting the fuhrman grade of clear-cell renal carcinoma. *Cancer Imaging* (2019) 19:1–7. doi: 10.1186/s40644-019-0195-7

7. Ljungberg B, Albiges L, Abu-Ghanem Y, Bensalah K, Dabestani S, Fernández-Pello S, et al. European Association of urology guidelines on renal cell carcinoma: the 2019 update. *Eur Urol* (2019) 75:799–810. doi: 10.1016/j.eururo.2019.02.011

8. Lane BR, Samplaski MK, Herts BR, Zhou M, Novick AC, Campbell SC. Renal mass biopsy–a renaissance? *J Urol* (2008) 179:20–7. doi: 10.1016/j.juro.2007.08.124

9. Cornelis F, Tricaud E, Lasserre A, Petitpierre F, Bernhard J, Le Bras Y, et al. Multiparametric magnetic resonance imaging for the differentiation of low and high grade clear cell renal carcinoma. *Eur Radiol* (2015) 25:24–31. doi: 10.1007/s00330-014-3380-x

10. Chen C, Kang Q, Xu B, Guo H, Wei Q, Wang T, et al. Differentiation of low-and high-grade clear cell renal cell carcinoma: Tumor size versus ct perfusion parameters. *Clin Imaging* (2017) 46:14–9. doi: 10.1016/j.clinimag.2017.06.010

11. Wu G, Zhao Z, Yao Q, Kong W, Xu J, Zhang J, et al. The study of clear cell renal cell carcinoma with mr diffusion kurtosis tensor imaging and its histopathologic correlation. *Acad Radiol* (2018) 25:430–8. doi: 10.1016/j.acra.2017.10.016

12. Lin F, Cui EM, Lei Y. Luo lp. ct-based machine learning model to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Abdominal Radiol* (2019) 44:2528–34. doi: 10.1007/s00261-019-01992-7

13. Lai S, Sun L, Wu J, Wei R, Luo S, Ding W, et al. Multiphase contrast-enhanced ct-based machine learning models to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Cancer Manage Res* (2021) 13:999–1008. doi: 10.2147/CMAR.S290327

14. Nikpanah M, Xu Z, Jin D, Farhadi F, Saboury B, Ball MW, et al. A deep-learning based artificial intelligence (ai) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic mri. *Clin Imaging* (2021) 77:291–8. doi: 10.1016/j.clinimag.2021.06.016

15. Lin F, Ma C, Xu J, Lei Y, Li Q, Lan Y, et al. A ct-based deep learning model for predicting the nuclear grade of clear cell renal cell carcinoma. *Eur J Radiol* (2020) 129:109079. doi: 10.1016/j.ejrad.2020.109079

16. Coy H, Hsieh K, Wu W, Nagarajan MB, Young JR, Douek ML, et al. Deep learning and radiomics: the utility of google tensorflow? inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic ct. *Abdominal Radiol* (2019) 44:2009–20. doi: 10.1007/s00261-019-01929-0

17. Han S, Hwang SI, Lee HJ. The classification of renal cancer in 3-phase ct images using a deep learning method. *J Digital Imaging* (2019) 32:638–43. doi: 10.1007/s10278-019-00230-2

18. Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* (2021) 11:1384. doi: 10.3390/diagnostics11081384

19. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: Medical Image

Computing and Computer Assisted Intervention, Strasbourg, France, Cham:. Springer. (2021) 12901:36–46. doi: 10.1007/978-3-030-87193-2_4

20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014) arXiv:1409.1556.

21. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vision* (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y

22. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions, in: *IEEE conference on computer vision and pattern recognition*. Boston, USA: Computer Vision Foundation / IEEE. (2015) pp. 1–9. doi: 10.1109/CVPR.2015.7298594

23. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. *arXiv preprint* (2018) arXiv:1803.07728.

24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale arXiv preprint arXiv:2010.11929. (2020).

25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, in: IEEE conference on computer vision and pattern recognition. Las Vegas, USA: Computer Vision Foundation / IEEE. (2016) pp. 770–8. doi: 10.1109/CVPR.2016.90

26. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks, in: *IEEE conference on computer vision and pattern recognition*. Hawaii, USA: Computer Vision Foundation / IEEE. (2017). pp. 4700–8. doi: 10.1109/CVPR.2017.243

27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision, in: *IEEE conference on computer vision and pattern recognition*. Las Vegas, USA: Computer Vision Foundation / IEEE. (2016) pp. 2818–26. doi: 10.1109/CVPR.2016.308

28. Hu J, Shen L, Sun G. Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, USA: Computer Vision Foundation / IEEE. (2018). pp. 7132–41. doi: 10.1109/CVPR.2018.00745

29. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: Computer Vision Foundation / IEEE. (2020) pp. 10428–36. doi: 10.1109/CVPR42600.2020.01044

30. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers, in: *2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada: IEEE. (2021) 32–42. doi: 10.1109/ICCV48922.2021.00010

31. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, et al. Conformer: Local features coupling global representations for visual recognition.,In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada. IEEE. (2021) pp. 357–66. doi: 10.1109/ICCV48922.2021.00042

32. Deng Y, Lu L, Aponte L, Angelidi AM, Novak V, Karniadakis GE, et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. *NPJ Digital Med* (2021) 4:1–13. doi: 10.1038/s41746-021-00480-x

33. Schwessinger R, Gosden M, Downes D, Brown RC, Oudelaar AM, Telenius J, et al. Deepc: predicting 3d genome folding using megabase-scale transfer learning. *Nat Methods* (2020) 17:1118–24. doi: 10.1038/s41592-020-0960-3

34. Yu J, Deng Y, Liu T, Zhou J, Jia X, Xiao T, et al. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat Commun* (2020) 11:1–10. doi: 10.1530/ey.18.3.15

35. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162

36. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *J Digital Imaging* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

37. Akin O, Elnajjar P, Heller M, Jarosz R, Erickson B, Kirk S, et al. Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [TCGA-KIRC] collection. *Cancer Imaging Arch* (2016). Website: https://wiki.cancerimagingarchive.net/display/Public/TCGA-KIRC

38. Coy H, Young JR, Douek ML, Pantuck A, Brown MS, Sayre J, et al. Association of qualitative and quantitative imaging features on multiphasic multidetector ct with tumor grade in clear cell renal cell carcinoma. *Abdominal Radiol* (2019) 44:180–9. doi: 10.1007/s00261-018-1688-8

39. Ding J, Xing Z, Jiang Z, Chen J, Pan L, Qiu J, et al. Ct-based radiomic model predicts high grade of clear cell renal cell carcinoma. *Eur J Radiol* (2018) 103:51–6. doi: 10.1016/j.ejrad.2018.04.013

40. Ishigami K, Leite LV, Pakalniskis MG, Lee DK, Holanda DG, Kuehn DM. Tumor grade of clear cell renal cell carcinoma assessed by contrast-enhanced computed tomography. *Springerplus* (2014) 3:1–7. doi: 10.1186/2193-1801-3-694

41. Han C, Rundo L, Murao K, Noguchi T, Shimahara Y, Milacski ZÁ, et al. Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *BMC Bioinf* (2021) 22:1–20. doi: 10.1186/s12859-020-03936-1

42. Wu Y, Hatipoglu S, Alonso-Álvarez D, Gatehouse P, Li B, Gao Y, et al. Fast and automated segmentation for the three-directional multi-slice cine myocardial velocity mapping. *Diagnostics* (2021) 11:346. doi: 10.3390/diagnostics11020346

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fonc.2022.958065

OPEN ACCESS

EDITED BY
Xiongbing Zu,
Xiangya Hospital, Central South University, China

REVIEWED BY
Tiezheng Qi,
Xiangya School of Medicine, Central South University, China
Zhe Min,
University College London, United Kingdom

*CORRESPONDENCE
Xiaoying Wang
wangxiaoying@bjmu.edu.cn

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 31 May 2022
ACCEPTED 12 September 2022
PUBLISHED 29 September 2022

CITATION
Zhu L, Gao G, Zhu Y, Han C, Liu X, Li D, Liu W, Wang X, Zhang J, Zhang X and Wang X (2022) Fully automated detection and localization of clinically significant prostate cancer on MR images using a cascaded convolutional neural network. *Front. Oncol.* 12:958065. doi: 10.3389/fonc.2022.958065

Fully automated detection and localization of clinically significant prostate cancer on MR images using a cascaded convolutional neural network

Lina Zhu[1], Ge Gao[2], Yi Zhu[3], Chao Han[2], Xiang Liu[2], Derun Li[4], Weipeng Liu[5], Xiangpeng Wang[5], Jingyuan Zhang[5], Xiaodong Zhang[2] and Xiaoying Wang[2*]

[1]Department of Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2]Department of Radiology, Peking University First Hospital, Beijing, China, [3]Department of Clinical & Technical Support, Philips Healthcare, Beijing, China, [4]Department of Urology, Peking University First Hospital, Beijing, China, [5]Department of Development and Research, Beijing Smart Tree Medical Technology Co. Ltd., Beijing, China

Purpose: To develop a cascaded deep learning model trained with apparent diffusion coefficient (ADC) and T2-weighted imaging (T2WI) for fully automated detection and localization of clinically significant prostate cancer (csPCa).

Methods: This retrospective study included 347 consecutive patients (235 csPCa, 112 non-csPCa) with high-quality prostate MRI data, which were randomly selected for training, validation, and testing. The ground truth was obtained using manual csPCa lesion segmentation, according to pathological results. The proposed cascaded model based on Res-UNet takes prostate MR images (T2WI+ADC or only ADC) as inputs and automatically segments the whole prostate gland, the anatomic zones, and the csPCa region step by step. The performance of the models was evaluated and compared with PI-RADS (version 2.1) assessment using sensitivity, specificity, accuracy, and Dice similarity coefficient (DSC) in the held-out test set.

Results: In the test set, the per-lesion sensitivity of the biparametric (ADC + T2WI) model, ADC model, and PI-RADS assessment were 95.5% (84/88), 94.3% (83/88), and 94.3% (83/88) respectively (all $p > 0.05$). Additionally, the mean DSC based on the csPCa lesions were 0.64 $\pm$ 0.24 and 0.66 $\pm$ 0.23 for the biparametric model and ADC model, respectively. The sensitivity, specificity, and accuracy of the biparametric model were 95.6% (108/113), 91.5% (665/727), and 92.0% (773/840) based on sextant, and were 98.6% (68/69), 64.8% (46/71), and 81.4% (114/140) based on patients. The biparametric model had a similar performance to PI-RADS assessment ($p > 0.05$) and had higher specificity than the ADC model (86.8% [631/727], $p< 0.001$) based on sextant.

Conclusion: The cascaded deep learning model trained with ADC and T2WI achieves good performance for automated csPCa detection and localization.

KEYWORDS

deep learning, prostatic neoplasms, magnetic resonance imaging, detection, localization

Introduction

Prostate cancer (PCa) is one of the most common malignant tumors in men worldwide. The clinical behavior of PCa ranges from low-grade indolent that is generally considered to be non-life-threatening to high-grade aggressive tumors with a Gleason of Score 7–10, i.e. clinically significant PCa (csPCa), that may progress rapidly to metastatic disease and may be life-threatening (1). Multiparametric magnetic resonance imaging (mpMRI) has adopted an increasingly significant role in the detection and localization of csPCa, as well as in guiding targeted biopsy (2). Recent large-scale clinical trials have demonstrated that the use of mpMRI before biopsy increases the detection of csPCa, while reducing the detection of those deemed clinically insignificant (3, 4). Furthermore, using mpMRI to triage male patients may enable a quarter to half of them to avoid the need for biopsy (3, 5). To standardize and improve the interpretation of prostate mpMRI, the use of the Prostate Imaging Reporting and Data System (PI-RADS) has been recommended and updated (2, 6). However, the interobserver agreement for subjective evaluation using PI-RADS (version 2) is moderate and influenced by the readers' experience (7, 8). Additionally, PI-RADS (version 2.1) has shown no significant improvements in overall diagnostic performance compared to PI-RADS (version 2.0) (9, 10). As there is arguably a trend in more people with clinically suspected csPCa undergoing prostate mpMRI, it is clinically desirable to develop more accurate and automated methods for prostate mpMRI interpretation.

In the recent years, artificial intelligence (AI) methods, particularly deep learning, have achieved promising results in automated csPCa diagnosis of mpMRI (11–14). A range of proposed deep learning algorithms were trained based on prior annotated regions of interest (ROIs) to classify them as cancerous or noncancerous lesions (11, 15), or slices, in which each individual image was classified as cancerous or not (16, 17). These methods were unable to precisely detect and locate csPCa, and such predicted results may not be directly applied to clinical practice. Some computer-aided diagnosis (CAD) systems developed for csPCa were based on the manual or semi-automatic segmentation of the prostate gland (13, 18), which also limits their direct clinical use. With the development of the deep convolutional neural network (CNN), some approaches for csPCa detection have been fully automatic with an area under the receiver operator characteristics curve of 0.75–0.86 (12, 19, 20). More studies are needed to improve and optimize these models. Although many generalized AI models have been developed, few studies have reported on how to integrate AI-based prediction into the clinical workflow. More explorations are demanded to move the prostate AI systems from the laboratory to the clinic with perfect output.

Our study aimed to develop a fully automated cascaded deep learning model for the detection and localization of csPCa using apparent diffusion coefficient (ADC) maps and T2-weighted imaging (T2WI), as well as to seamlessly integrate these AI predictions into the radiological workflow using the structured report.

Materials and methods

Study subjects

Our institute's review board approved this retrospective study and waived the need for informed consent. The inclusion criteria for the study were mpMRI scans performed on a GE 750 3.0T MRI scanner at Peking University First Hospital from March 2017 to February 2020 on consecutive patients who underwent mpMRI before a biopsy, with a clinical suspicion of PCa due to an elevated serum prostate-specific antigen (PSA) level, abnormal digital rectal examination (DRE), and/or abnormal transrectal ultrasound (TRUS) results. Exclusion criteria were patients without a subsequent biopsy performed within 3 months after mpMRI examination, a negative biopsy for csPCa without clinical follow-up >1 year, or showing potential csPCa during the clinical follow-up (progression of PSA or MR findings), as well as images with severe artifacts or incomplete pathology results which could not be matched with MR images. In total, 347 patients were included. Figure 1 displays the flow diagram for the inclusion of patients in the study. In this study, csPCa was defined as the International Society of Urological Pathology Gleason grade group ≥2, i.e., Gleason Score ≥7. The characteristics of the 235 patients with csPCa included are shown in Table 1. The other 112 patients without csPCa (labeled non-csPCa with a mean age

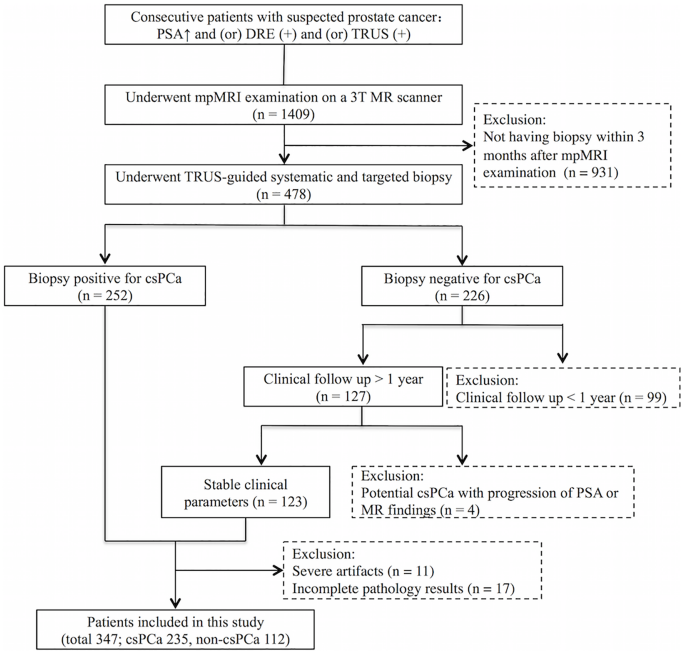


FIGURE 1
Flow diagram for inclusion of patients into the study.

of 64.1 ± 7.5 years) had a median PSA level of 8.0 ng/ml, with an interquartile range of 6.6–13.1 ng/ml. The patients were randomly selected to populate the datasets for training (145 csPCa, 35 non-csPCa), validation (21 csPCa, 6 non-csPCa), and testing (69 csPCa, 71 non-csPCa).

MRI sequences

All of the mpMRI examinations were performed on a 3-T MR machine (Discovery MR750, GE Medical Systems). A 32-channel abdominal phased array coil was used as the receiving coil. All patients were scanned following the unified prostate mpMRI protocol. The main sequence parameters are summarized in Table 2. The ADC map was automatically generated by the MR vendor software based on diffusion-weighted imaging (DWI) data with different b values. Concerning the patients included in this study, anonymized images were exported in the Digital Imaging and Communication in Medicine (DICOM) format.

PI-RADS assessment

All of the mpMRI cases in the dataset of the test were retrospectively interpreted according to PI-RADS (version 2.1) by a urogenital radiologist with 10 years of experience in prostate MRI diagnosis. The radiologist was informed of the clinical information of the patients, such as age, biopsy history, PSA, etc., but was blinded to the pathology results and the previous MRI reports. The lesions detected were delineated on a prostate sector map.

Reference standard and annotation

All of the patients in this study underwent TRUS-guided systematic and targeted biopsy using 12- or six-core needles. For cognitive targeting, lesions suspected of malignancy on mpMRI had been marked on a prostate sector map (6) using structured reports by five dedicated urogenital radiologists during the clinical routine. Before the biopsy, MR images would be reviewed by at least one urogenital radiologist and one urologist at a multidisciplinary meeting to ensure the accuracy of suspicious lesions localization. The urologists obtained additional needle cores (two- to five-core needles) for each of the suspicious lesions during the TRUS-guided biopsy. Histopathology analysis of each specimen was performed by a urogenital pathologist with 11 years of experience.

Two experienced urogenital radiologists (X and Y with 7 and 13 years of experience in prostate MRI diagnosis, respectively) retrospectively reviewed all csPCa cases and mapped the detailed

TABLE 1 Characteristics of patients with csPCa.

Characteristics	Patients with csPCa (n = 235)
Mean age (years)	70.2 ± 8.6
Median PSA (ng/mL)	16.3 (9.7–32.6)
Per-patient maximum Gleason score	
3 + 4	76
4 + 3	59
3 + 5, 5 + 3, 4 + 4	46
4 + 5, 5 + 4	54
No. of csPCa lesions per patient	
One lesion	173
Two lesions	49
Three lesions	10
Four lesions	3
Zone distribution of csPCa lesions	
Peripheral zone	212
Transition zone	101

csPCa, clinically significant prostate cancer; PSA, serum prostate-specific antigen.

pathology results of the csPCa foci to the MR images with consensus. The ground truth of the csPCa lesion segmentation was obtained using manual delineation by the urogenital radiologist (X), in consensus with and under the supervision of the other urogenital radiologist (Y), using the open-source segmentation software ITK-SNAP (version 3.6 2015; available at www.itksnap.org) (21). The format of ADC and T2WI was converted from DICOM to NIFTI. Three-dimensional volumes of interest (VOIs) were manually drawn along the boundaries of the csPCa lesions on consecutive axial sections of ADC images.

Image preprocessing

After collecting the mpMRI data, the first step of image preprocessing is T2WI and DWI/ADC image registration. Patient motion is minimal and the two sequences are temporally close to each other during the scanning. T2WI and DWI/ADC images were registered *via* rigid transformation using the coordinate information stored in the DICOM image headers. B-spline interpolation to the third order was employed for all MR image interpolation tasks, while Gaussian label interpolation was used for the csPCa and prostate segmentation masks. Following this, a coarse segmentation of the prostate was obtained by K-means clustering in DWI high b value images to localize the prostate region. Once the prostate region was identified, the images were cropped to a patch of size 128 × 128. The prostate region of interest was then normalized into the range of [0, 1]. We augmented the data in the training set by mirroring, random rotation (rotation angle within 10°), and adding noise (within 0.001, which means each pixel value randomly fluctuates within one thousandth).

Deep learning framework

The base architecture for the deep learning framework used in this study is a CNN inspired by the 2D U-Net (22) and Res-Net (23) architectures and is termed Res-UNet. U-Net is one of the end-to-end methods of deep learning, which is also a pixel-to-pixel method and, with long skip connections, considers feature maps of the encoder path to obtain good segmentation performance in medical images. Res-Net (23) proposed a residual connection architecture to make the network deeper and avoid gradient vanishing and take advantage of strengths from both architectures by modifying the original U-Net architecture and adding residual blocks into the contracting and symmetric expanding paths of the U-Net architecture. In building the Res-UNet, we define a basic convolution operator by a 3 × 3 convolution (Conv) followed by a batch normalization (BN) and a rectified linear unit (ReLu). The residual block was designed by using a 1 × 1 Conv layer, plus an addition operation and ReLU function. Figure 2 depicts the Res-UNet architecture.

Following McNeal's criterion (24), the prostate is typically partitioned into two distinct zones: the Central Gland (CG, including both the transition zone and the central zone) and the Peripheral Zone (PZ). PCa lesions vary in frequency and malignancy depending on the zone, and there are different evaluation criteria for different regions in the PI-RADS.

TABLE 2 The main sequence parameters in this study.

	T2WI	DWI	DCE
Field of view (mm)	240 × 240	240 × 240	260 × 260
Acquisition Matrix	320 × 256	96 × 96	320 × 192
Repetition time (ms)	3200	3000	4
Echo time (ms)	85	60	1.3
Flip angle (degrees)	111	90	13
Slice thickness (mm)	4	4	3
Additional information	…	b values: 0–1400 s/mm^2	Temporal resolution = 13 s;18 phases

T2WI, T2-weighted imaging; DWI, diffusion-weighted imaging; DCE, dynamic contrast material enhancement.

Therefore, just like a radiologist, a model for automated PCa detection and classification will invariably benefit from having both CG and PZ mask priors provided as inputs, in addition to the mpMRI. Accordingly, a cascading system of a three-segmentation Res-UNet which was previously developed in our institution (25) was developed for automatic prostate CG and PZ segmentation and PCa lesion segmentation. The cascade is designed to decompose the multi-task segmentation problem into a sequence of three smaller binary segmentation problems according to the subregion hierarchy. As can be seen in Figure 2, the first Res-UNet model takes prostate MR images (T2WI + ADC or only ADC) as inputs and produces a prostate segmentation mask as an output. Then, the second model takes the MR images and the prostate segmentation mask which were obtained in the previous step as inputs and produces a PZ segmentation mask. The CG segmentation mask can be computed by subtracting the PZ mask from the prostate mask. Finally, the last model takes the MR images, the PZ mask, and the CG mask as inputs and segments to the csPCa.

Training setup

All of the training steps were performed using a GPU NVIDIA Tesla P100 16G at Peking University First Hospital, using the operating system Ubuntu 16.04. The software and packages used included Python 3.6, Pytorch 0.4.1, Opencv 3.4.0.12, Numpy 1.16.2, and SimpleITK 1.2.0. The input data were 128 × 128 images of ADC alone and ADC combined with T2WI, respectively, with an annotation of the csPCa lesions. The automated segmentation of the whole prostate gland and its different zones was completed using the previously developed and described method (25). For training the architectures for csPCa segmentation, the batch size was set as 20 with a learning rate of 0.0001. The networks were trained for 120 epochs. The pixel classification threshold was 0.5, while Adam was used as a training optimizer. The Dice similarity coefficient (DSC) was used to evaluate the performance of the networks in the segmentation of the csPCa, which is calculated as

$$\mathrm{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}$$

here X is the pixel set of csPCa segmented manually as the ground truth and Y is the pixel set of csPCa prediction by the model.

Prediction results integrated into the structured report

Initially, the prediction results were "csPCa" or "non-csPCa" depending on the patient concerned. When "csPCa" was the output, the three-dimensional diameter of the suspicious csPCa lesions and the whole prostate gland would be filled into the structured report, as well as the key image of the suspicious csPCa lesions (Supplementary Figure S1). Furthermore, sextant localization of the suspicious csPCa lesions would be labeled in the prostate sector map. The prostate sextant is defined according to the standard sextant biopsy (26), based on the automatic segmentation of the prostate gland.

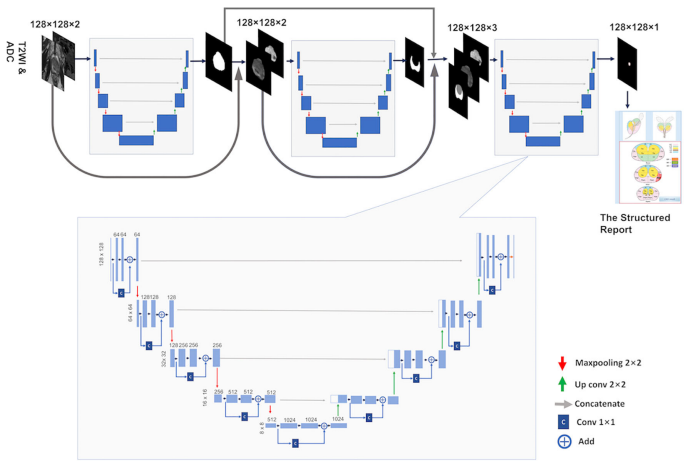


FIGURE 2
The briefarchitecture of the proposed weighted Res-Unet and the overall pipeline of our approach.

Statistical analysis

Statistical analysis was carried out using SPSS 20.0 and MedCalc 15.8. We evaluated the performance of the biparametric (ADC + T2WI) model, ADC model, and PI-RADS assessment using the testing set. For PI-RADS assessment, PI-RADS ≥3 was considered positive for csPCa. For per-lesion analysis, to limit the influence of very small overlap regions, only the predicted lesions of the model overlapping ≥50% of the manual csPCa segmentation lesions in at least one slice were considered as true positive. Otherwise, the predicted lesions were considered to be false positives. For sextants analysis, only sextants overlapping at least 50% of an MRI lesion, or being occupied at least 50% by an MRI lesion, were considered to contain the MRI lesion (13). For per-patient analysis, if a patient had one or more than one csPCa lesions, the prediction of the model or PI-RADS assessment, was considered as true positive when at least one csPCa lesion was detected. On the other hand, for a patient without csPCa, the prediction was considered as false positive as long as one lesion was predicted. The performance of the models and PI-RADS assessment for csPCa detection and localization were evaluated based on the lesions, sextants, and patients, respectively. The sensitivity, specificity, and accuracy of the models and PI-RADS were evaluated and compared using the McNemar test. A p value of less than 0.05 was considered statistically significant.

Results

Based on lesions

In the test set, 88 csPCa lesions were included and the mean greatest dimension was 1.6 ± 0.7 cm. Table 3 depicts the sensitivity of the models and PI-RADS assessment on the per-lesion analysis. The per-lesion sensitivity of the biparametric model, ADC model, and PI-RADS assessment was 95.5% (84/88), 94.3% (83/88), and 94.3% (83/88), respectively, with all p > 0.05. For the csPCa lesions with the greatest dimension ≥1.5cm, the sensitivity of the biparametric model was 100%, and the sensitivity of the ADC model and PI-RADS assessment was 97.6% (40/41, with p > 0.05). The sensitivity showed no significant difference between the models and PI-RADS regardless of whether the csPCa lesions were in the PZ or the TZ.

In addition, the mean DSC based on csPCa lesions in the test were 0.64 ± 0.24 and 0.66 ± 0.23 for the biparametric model and the ADC model, respectively. Figure 3 demonstrates examples of the csPCa segmentation of the biparametric model.

Based on sextants

A total of 840 sextants from the test set were analyzed, including 113 sextants of csPCa and 727 sextants of non-csPCa. The diagnostic efficacy and comparisons of the models and the PI-RADS assessment based on sextants are summarized in Tables 4 and 5. The biparametric model and PI-RADS assessment both had relatively high sensitivity, specificity and accuracy, i.e., 95.6% (108/113) *vs.* 92.9% (105/113), 91.5% (665/727) *vs.* 92.2% (670/727), and 92.0% (773/840) *vs.* 92.3% (775/840), respectively, all with p > 0.05. The ADC model demonstrated a comparable sensitivity of 91.2% (103/113) when compared with the biparametric model where p = 0.125, while the ADC model had a specificity of 86.8% (631/727) and accuracy of 87.4% (734/840), which were both lower than the biparametric model (all p< 0.001). Considering all the mpMRI sequences and detailed pathological results, 61.5% (59/96) of the false-positive sextants from the ADC model were hyperplastic nodules and asymmetric central zone, which is shown in Figure 4, while the ratio was 45.2% (28/62) for the biparametric model.

Based on patients

For the 140 patients (csPCa, 69; non-csPCa, 71) in the test set, the performance and comparisons of the models and the PI-RADS assessment based on patients are shown in Tables 4 and 5. Biparametric model and PI-RADS assessment had comparable per-patient sensitivity, specificity, and accuracy, i.e. 98.6% (68/69) *vs.* 98.6% (68/69), 64.8% (46/71) *vs.* 66.2% (47/71) and 81.4% (114/140) *vs.* 82.1% (115/140), respectively (all p > 0.05). ADC model had a similar sensitivity of 97.1% (67/69) compared with the combined model and PI-RADS assessment. The specificity

TABLE 3 Per-lesion sensitivity of the models and PI-RADS assessment.

	Biparametric model (%)	ADC model (%)	PI-RADS (%)	p
Total	95.5 (84/88) [88.8, 98.8]	94.3 (83/88) [87.2, 98.1]	94.3 (83/88) [87.2, 98.1]	>0.05
Peripheral zone	95.0 (57/60) [86.1, 99.0]	93.3 (56/60) [83.8, 98.2]	96.7 (58/60) [88.5, 99.6]	>0.05
Transition zone	96.4 (27/28) [81.7, 99.9]	96.4 (27/28) [81.7, 99.9]	89.3 (25/28) [71.8, 97.7]	>0.05
Dimension 0.4-1.5cm	91.5 (43/47) [79.6, 97.6]	91.5 (43/47) [79.6, 97.6]	91.5 (43/47) [79.6, 97.6]	>0.05
Dimension ≥1.5cm	100.0 (41/41) [96.4, 100]	97.6 (40/41) [87.1, 99.9]	97.6 (40/41) [87.1, 99.9]	>0.05

PI-RADS, Prostate Imaging Reporting and Data System.

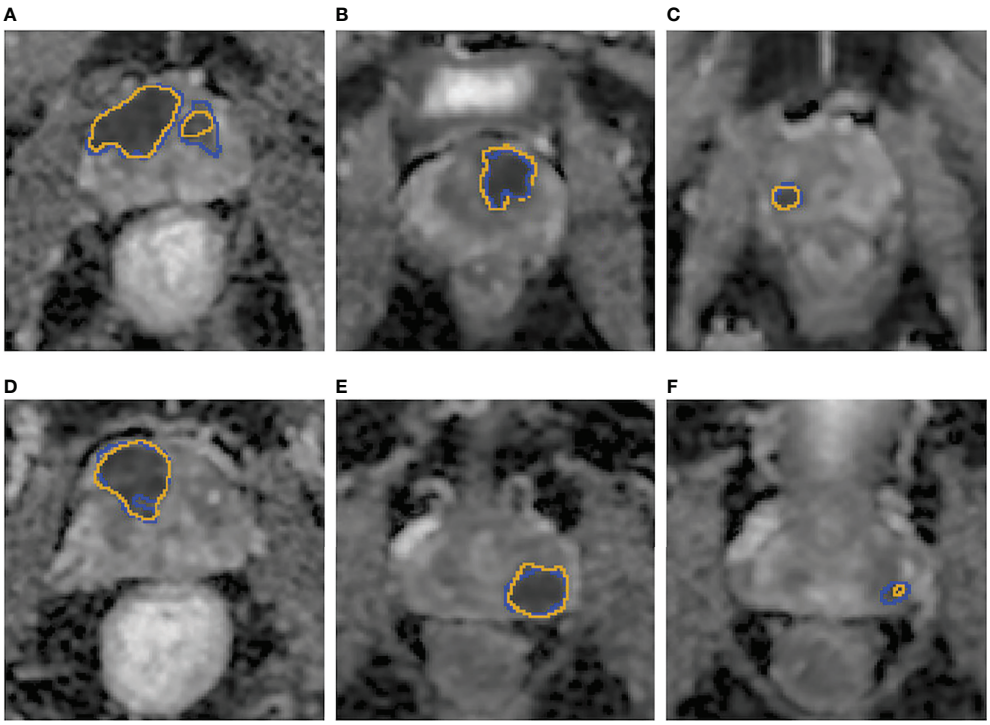


FIGURE 3
(A–F) Examples of the csPCa lesion segmentation performance of the biparametric model. The prediction results **(A–F)**, yellow line on the ADC map were highly consistent with the manual annotation **(A–F)**, blue line on ADC map by experienced urogenital radiologists according to pathological results.

and accuracy of the biparametric model were slightly higher than those of the ADC model (specificity, 54.9%; accuracy, 75.7%); however, the statistical significance was not reached (p = 0.118, 0.077).

Discussion

Our approach using cascaded CNNs could automatically detect and segment the suspicious csPCa lesions on MR images without any human intervention. The whole prediction process could be completed within a few seconds per case, which was much faster than human interpretation using PI-RADS, which normally takes several minutes. There are several benefits of using a cascaded framework for the segmentation of hierarchical structures. First, many proposed methods try to solve the segmentation problem using a single neural network. Considering the great variability in the shape, size, texture, and appearance of the prostate gland and PCa, we suggest using cascaded CNNs for the segmentation task to improve

TABLE 4 Performance of the models and PI-RADS assessment based on sextants and patients.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Based on sextant			
Biparametric model	95.6 (108/113) [90.0, 98.6]	91.5 (665/727) [89.2, 93.4]	92.0 (773/840)
ADC model	91.2 (103/113) [84.3, 95.7]	86.8 (631/727) [84.1, 89.2]	87.4 (734/840)
PI-RADS	92.9 (105/113) [86.5, 96.9]	92.2 (670/727) [90.0, 94.0]	92.3 (775/840)
Based on patient			
Biparametric model	98.6 (68/69) [92.2, 99.9]	64.8 (46/71) [52.5, 75.8]	81.4 (114/140)
ADC model	97.1 (67/69) [89.9, 99.7]	54.9 (39/71) [42.7, 66.8]	75.7 (106/140)
PI-RADS	98.6 (68/69) [92.2, 99.9]	66.2 (47/71) [54.0, 77.0]	82.1 (115/140)

PI-RADS, Prostate Imaging Reporting and Data System.
Data in brackets are 95% CIs.

TABLE 5 Comparisons of the models and PI-RADS assessment based on sextants and patients.

	Sensitivity	Specificity	Accuracy
Based on sextants			
Biparametric model *vs.* ADC model	0.125	**<0.001**	**<0.001**
Biparametric model *vs.* PI-RADS	0.508	0.630	0.910
ADC model *vs.* PI-RADS	0.754	**<0.001**	**<0.001**
Based on patients			
Biparametric model *vs.* ADC model	1.000	0.118	0.077
Biparametric model *vs.* PI-RADS	1.000	1.000	1.000
ADC model *vs.* PI-RADS	1.000	0.077	0.064

PI-RADS, Prostate Imaging Reporting and Data System.
Bold characters indicate that the difference was statistically significant ($p < 0.05$).

the segmentation accuracy, and each network can focus on one segmentation problem. Thus, they are easier to train and can reduce over-fitting. Second, in consideration of the PCa lesions, these can vary in frequency and malignancy depending on the zone; the hierarchical pipeline follows the anatomical structures of the prostate and uses them as spatial constraints. Thus, the model for automated csPCa detection and classification will likely benefit.

The results demonstrated that the biparametric model had high sensitivity (95.5%, 95.6%, and 98.6% based on lesions, sextants, and patients respectively) and acceptable specificity (64.8%, 91.5%; based on patients and sextants) and had comparable performance to PI-RADS evaluation by an experienced radiologist, which is consistent with Schelb's findings (13). The preliminary results of our study add to the evidence that fully automated deep learning models for csPCa

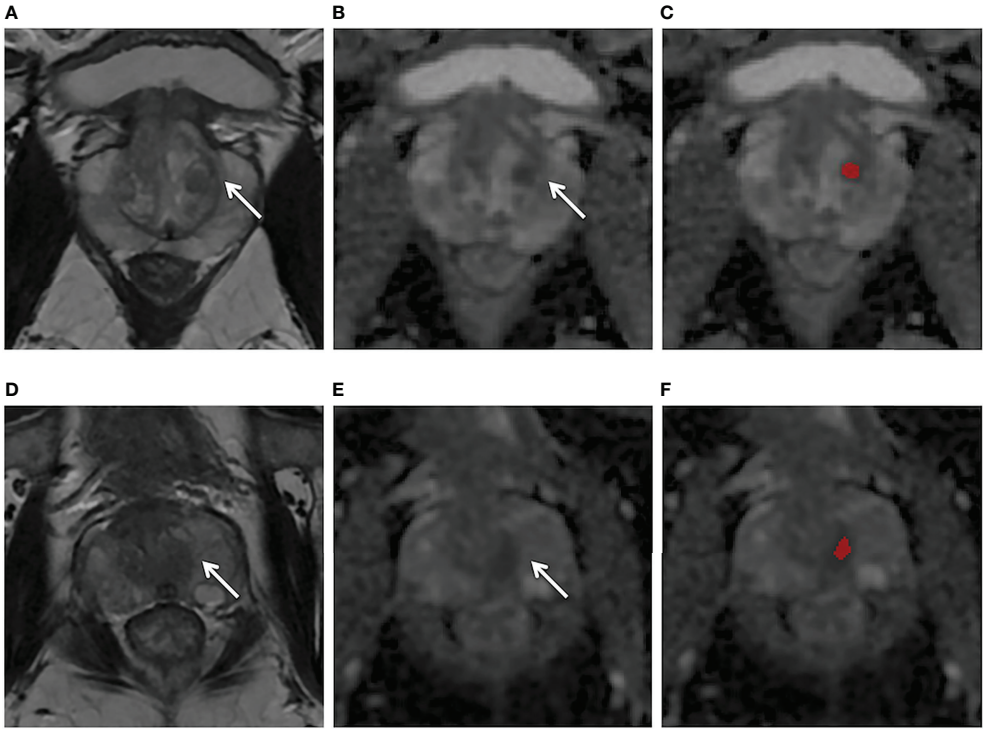


FIGURE 4
(A–C) Axial MR images obtained in a 56-year-old patient with a PSA level of 4.2 ng/ml and with negative biopsy findings. T2WI **(A)** showed a heterogeneous encapsulated nodule in the left transition zone (arrow) and the ADC map **(B)** showed hypointensity (arrow). The ADC model **(C)** appeared false positive (red region). **(D–F)** Axial MR images obtained in a 64-year-old patient with a PSA level of 5.9 ng/ml and with negative biopsy findings. T2W **(D)** and ADC **(E)** showed a normal left central zone, while the ADC model **(F)** appeared false positive in this area (red mark). The biparametric model gave negative predictive values for both cases.

detection have now even reached the level of an experienced radiologist (13, 27). Further prospective studies based on large consecutive data are needed for clinical validation. Furthermore, our model could also determine the boundary of csPCa precisely. The DSC based on csPCa lesions was 0.64 and 0.66 for the biparametric model and ADC model, respectively, which was higher than that in similar studies on csPCa detection that reported 0.35–0.58 (13, 28, 29). The good segmentation performance would facilitate the 3D prostate MRI-TRUS fusion targeted biopsy. Additionally, a 3D model for the visualization of csPCa and the adjacent vital structures, based on accurate segmentation, may be helpful for the urologist in the surgery, as well as for patient education (30). However, DSC as a voxel-level metric remains limited for lesion-level PCa detection and can misrepresent the accuracy for evaluating the localization of multifocal PCa (31, 32). Therefore, our study based on the actual clinical practice mainly used sensitivity and specificity at the lesion, sextant, and patient levels to comprehensively evaluate the performance of the model.

T2WI and ADC derived from DWI are recommended by PI-RADS as the most important sequences for the evaluation of TZ and PZ lesions, respectively. Many studies demonstrated that the diagnostic performance of biparametric MRI without DCE was similar to those of mpMRI (33, 34). Therefore, this study mainly used those two parameters to develop the model. Additionally, the ADC of DWI is considered to be the current best monoparametric sequence of prostate MRI assessment, which is reported to have a strong relationship with the Gleason scores (GS) of PCa and is even superior to TRUS-guided prostate biopsy for the assessment of PCa aggressiveness (35–37). For this reason, the proposed model trained with monoparametric ADC was evaluated specifically. The results of our study demonstrated that the monoparameter ADC model had a high sensitivity for csPCa detection (94.3% [83/88] and 97.1% [67/69], based on lesions and patients, respectively), which showed no significant difference ($p > 0.05$) with the biparametric model and PI-RADS assessment, regardless of whether the csPCa lesions were located on the PZ or TZ. Zabihollahy et al.'s study also showed that deep learning using only ADC was highly sensitive and could even reach a 100% sensitivity at the level of dominant PZ tumor detection (12), which is slightly higher than ours. The reason may be that their study only considered the most dominant lesion on PZ, which was easily identifiable, while our study detected all the csPCa on the MR images. Further research with larger volumes of testing data is needed to verify the performance of the ADC model. The specificity of the biparametric model in our study was higher than that of the ADC model based on sextants ($p< 0.001$). When compared with the ADC model, the biparametric model outputted fewer false-positive lesions such as hyperplastic nodules and the central zone. Nevertheless, the high sensitivity of the ADC model using a single parameter instead of time-consuming mpMRI may facilitate the promotion of prostate MRI screening.

In contrast with studies that were training models using public data (15, 38), the data in our study were collected consecutively based on real-world clinical scenarios, which would allow the model to be more easily integrated into a clinical setting. The amount of data in the training and validation sets (145 and 21 csPCa cases, including 1204 slices and 173 slices, respectively, after automatic prostate segmentation) was comparable to that used in some studies (11, 12), and was larger than that in other studies (19, 20). Therefore, we put more data in the testing set (69 csPCa, 71 non-csPCa cases) to better evaluate and verify the generalization ability of the model. Our proposed biparametric model yielded high sensitivity of 98.6% for csPCa detection based on patients, as well as other studies with the sensitivity ranging from 82.9% to 97% (13, 14, 39–41). However, it is worth noting that the specificity was not as high as expected in our study (64.8%) and in other studies (47%–76%) (12–14, 39). That is to say, the success of the AI came at the cost of a high false-positive rate of even 50% (42). Yu et al. (43)proposed a cascaded approach to reduce the false positive for PCa detection, where the second-stage classifier was able to reduce false positives at the expense of nearly an 8% decrease in detection sensitivity. Saha et al. (29) present a multi-stage 3D CAD model for csPCa localization in biparametric MRI with the addition of a residual patch-wise 3D classifier to improve the model specificity. The results demonstrated that up to 12.89% less false positives were generated per patient, while retaining the same sensitivity (92.29%) as before. Min et al. (44) explored the feasibility of controlling the false positives/negatives during training by incorporating the cost-sensitive classification losses. More studies are needed to further explore how to improve the specificity of the prostate CAD.

An optimal AI model should not only have good performance for csPCa diagnosis but should also have a perfect output form facilitating clinical practice. One advantage of the proposed approach is that it had a perfect output and added AI into the radiological workflow seamlessly by automatically integrating the prediction results into structured reports, which makes this approach more convenient for clinical application. At present, our model can output whether a patient has csPCa or not and the size of the prostate gland. If csPCa lesions were found, it would further output the size of csPCa lesions and also mark the area of csPCa on MR images. These prediction results could be automatically transferred into a structured report before radiologists open the reporting system. However, a complete structured report includes many other contents, such as whether the csPCa lesions invade the prostate capsule and adjacent structures, as well as lymph node condition, bone metastasis, etc. Our institution is now exploring each of the above, and some of them have achieved good performance (45). Our ultimate goal was to develop a fully automatic intelligent structured report, thereby freeing radiologists from heavy clinical paperwork. When radiologists open structured reports, they just need to check the accuracy of each item.

Several limitations of our study were as follows. First, all of the images were from a single MR machine in a single institution. Multi-center and multi-machine data functionality should be added to improve the generalization ability of the model in further studies. Secondly, this U-Net model was trained only using ADC and T2WI. Future research involving the addition of more MRI sequences and/or clinical information may be investigated to improve the performance of the model. On the other hand, the model in this study achieved good results with biparametric MRI, so streamlining MRI sequences with an advanced algorithm may be another possible research direction. Thirdly, even though the reference standard using TRUS-guided systematic and targeted biopsy had high sensitivity for csPCa, it still has a false-negative rate when compared with radical prostatectomy. Nevertheless, our cohorts may be optimal, for radical prostatectomy cohorts would exclude many patients who only had a prostate biopsy and could lead to bias. Finally, this model was only applicable to the detection and localization of csPCa instead of staging and active surveillance.

Conclusion

In conclusion, our study demonstrated that a cascaded deep learning model trained with ADC and T2WI achieved good performance for the fully automated detection and segmentation of csPCa and demonstrated comparable performance with an experienced radiologist using PI-RADS (version 2.1). The proposed approach can automatically integrate prediction results into the radiological workflow seamlessly by using a structured report. As a preliminary exploration, this study provided a reference for future AI clinical implementation. Further studies are needed to explore the optimal paradigm of AI clinical integration.

Data availability statement

The datasets presented in this article are not readily available because the datasets are privately owned by Peking University First Hospital and are not made public. Requests to access the datasets should be directed to XYW, wangxiaoying@bjmu.edu.cn.

Ethics statement

The studies involving human participants were reviewed and approved by Peking University First Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

LZ and XYW designed the study. CH, XL, and DL contributed to acquisition of data. LZ and XYW annotated the images data. YZ, WL, XPW, and JZ designed the model and implemented the main algorithm. LZ and GG analyzed the data. LZ wrote the paper. XYW and XZ reviewed the paper. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors would like to acknowledge Xin Yue and Suisui Zhang from the Beijing Smart Tree Medical Technology Co., Ltd., for their help in constructing the structured report.

Conflict of interest

Authors WL, XPW and JZ are employed by Beijing Smart Tree Medical Technology Co. Ltd.

The remaining 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.958065/full#supplementary-material

References

1. Johnson LM, Turkbey B, Figg WD, Choyke PL. Multiparametric MRI in prostate cancer management. *Nat Rev Clin Oncol* (2014) 11:346–53. doi: 10.1038/nrclinonc.2014.69

2. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur Urol* (2019) 76:340–51. doi: 10.1016/j.eururo.2019.02.033

3. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* (2017) 389:815–22. doi: 10.1016/S0140-6736(16)32401-1

4. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-Targeted or standard biopsy for prostate-cancer diagnosis. *N Engl J Med* (2018) 378:1767–77. doi: 10.1056/NEJMoa1801993

5. van der Leest M, Cornel E, Israel B, Hendriks R, Padhani AR, Hoogenboom M, et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naive men with elevated prostate-specific antigen: A Large prospective multicenter clinical study. *Eur Urol* (2019) 75:570–8. doi: 10.1016/j.eururo.2018.11.023

6. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, et al. PI-RADS prostate imaging - reporting and data system: 2015, version 2. *Eur Urol* (2016) 69:16–40. doi: 10.1016/j.eururo.2015.08.052

7. Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, et al. Interobserver reproducibility of the PI-RADS version 2 lexicon: A multicenter study of six experienced prostate radiologists. *Radiology* (2016) 280:793–804. doi: 10.1148/radiol.2016152542

8. Byun J, Park KJ, Kim MH, Kim JK. Direct comparison of PI-RADS version 2 and 2.1 in transition zone lesions for detection of prostate cancer: Preliminary experience. *J Magn Reson Imaging* (2020) 52:577–86. doi: 10.1002/jmri.27080

9. Rudolph MM, Baur ADJ, Cash H, Haas M, Mahjoub S, Hartenstein A, et al. Diagnostic performance of PI-RADS version 2.1 compared to version 2.0 for detection of peripheral and transition zone prostate cancer. *Sci Rep* (2020) 10:15982. doi: 10.1038/s41598-020-72544-z

10. Linhares Moreira AS, De Visschere P, Van Praet C, Villeirs G. How does PI-RADS v2.1 impact patient classification? a head-to-head comparison between PI-RADS v2.0 and v2.1. *Acta Radiol* (2021) 62:839–47. doi: 10.1177/0284185120941831

11. Aldoj N, Lukas S, Dewey M, Penzkofer T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *Eur Radiol* (2020) 30:1243–53. doi: 10.1007/s00330-019-06417-z

12. Zabihollahy F, Ukwatta E, Krishna S, Schieda N. Fully automated localization of prostate peripheral zone tumors on apparent diffusion coefficient map MR images using an ensemble learning method. *J Magn Reson Imaging* (2020) 51:1223–34. doi: 10.1002/jmri.26913

13. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of cancer at prostate MRI: Deep learning versus clinical PI-RADS assessment. *Radiology* (2019) 293:607–17. doi: 10.1148/radiol.2019190938

14. Cao R, Mohammadian Bajgiran A, Afshari Mirak S, Shakeri S, Zhong X, Enzmann D, et al. Joint prostate cancer detection and Gleason score prediction in mp-MRI *via* FocalNet. *IEEE Trans Med Imaging* (2019) 38:2496–506. doi: 10.1109/TMI.2019.2901928

15. Song Y, Zhang YD, Yan X, Liu H, Zhou M, Hu B, et al. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *J Magn Reson Imaging* (2018) 48:1570–7. doi: 10.1002/jmri.26047

16. Yoo S, Gujrathi I, Haider MA, Khalvati F. Prostate cancer detection using deep convolutional neural networks. *Sci Rep* (2019) 9:19518. doi: 10.1038/s41598-019-55972-4

17. Wang Z, Liu C, Cheng D, Wang L, Yang X, Cheng KT. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-End deep neural network. *IEEE Trans Med Imaging* (2018) 37:1127–39. doi: 10.1109/TMI.2017.2789181

18. Lemaitre G, Marti R, Rastgoo M, Meriaudeau F. Computer-aided detection for prostate cancer detection based on multi-parametric magnetic resonance imaging. *Annu Int Conf IEEE Eng Med Biol Soc* (2017) 2017:3138–41. doi: 10.1109/EMBC.2017.8037522

19. Seetharaman A, Bhattacharya I, Chen LC, Kunder CA, Shao W, Soerensen SJC, et al. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Med Phys* (2021) 48:2960–72. doi: 10.1002/mp.14855

20. Sanyal J, Banerjee I, Hahn L, Rubin D. An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence. *AMIA Jt Summits Transl Sci Proc* (2020) 2020:552–60.

21. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015

22. Ronneberger O, Fischer P, Brox T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells, W, Frangi A. (eds). in: *Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science*. Cham: Springer (2015) 9351:234–41. doi: 10.1007/978-3-319-24574-4_28

23. Nah S, Kim TH, Lee KM. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3883–91. doi: 10.48550/arXiv.1612.02177

24. McNeal JE. Anatomy of the prostate: an historical survey of divergent views. *Prostate* (1980) 1:3–13. doi: 10.1002/pros.2990010103

25. Zhu Y, Wei R, Gao G, Ding L, Zhang X, Wang X, et al. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J Magn Reson Imaging* (2019) 49:1149–56. doi: 10.1002/jmri.26337

26. Presti JC. Prostate biopsy: current status and limitations. *Rev Urol* (2007) 9:93–8.

27. Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *Eur Radiol* (2022) 32:2224–34. doi: 10.1007/s00330-021-08320-y

28. Vente C, Vos P, Hosseinzadeh M, Pluim J, Veta M. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Trans BioMed Eng* (2021) 68:374–83. doi: 10.1109/TBME.2020.2993528

29. Saha A, Hosseinzadeh M, Huisman H. End-to-end prostate cancer detection in bpMRI *via* 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med Image Anal* (2021) 73:102155. doi: 10.1016/j.media.2021.102155

30. Wake N, Rosenkrantz AB, Huang R, Park KU, Wysock JS, Taneja SS, et al. Patient-specific 3D printed and augmented reality kidney and prostate cancer models: impact on patient education. *3D Print Med* (2019) 5:4. doi: 10.1186/s41205-019-0041-3

31. Carass A, Roy S, Gherman A, Reinhold JC, Jesson A, Arbel T, et al. Evaluating white matter lesion segmentations with refined sørensen-dice analysis. *Sci Rep* (2020) 10:8242. doi: 10.1038/s41598-020-64803-w

32. Yan W, Yang Q, Syer T, Min Z, Punwani S, Emberton M, et al. The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation. *arXiv preprint arXiv* (2022) 2203:16415. doi: 10.48550/arXiv.2203.16415

33. Choi MH, Kim CK, Lee YJ, Jung SE. Prebiopsy biparametric MRI for clinically significant prostate cancer detection with PI-RADS version 2: A multicenter study. *AJR Am J roentgenol* (2019) 212:839–46. doi: 10.2214/AJR.18.20498

34. Tamada T, Kido A, Yamamoto A, Takeuchi M, Miyaji Y, Moriya T, et al. Comparison of biparametric and multiparametric MRI for clinically significant prostate cancer detection with PI-RADS version 2.1. *J Magn Reson Imaging* (2021) 53:283–91. doi: 10.1002/jmri.27283

35. Bonekamp D, Kohl S, Wiesenfarth M, Schelb P, Radtke JP, Gotz M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: Comparison to ADC values. *Radiology* (2018) 289:128–37. doi: 10.1148/radiol.2018173064

36. Kim TH, Kim CK, Park BK, Jeon HG, Jeong BC, Seo SI, et al. Relationship between Gleason score and apparent diffusion coefficients of diffusion-weighted magnetic resonance imaging in prostate cancer patients. *Can Urol Assoc J Nov-Dec* (2016) 10:E377–82. doi: 10.5489/cuaj.3896

37. Li C, Chen M, Wang J, Wang X, Zhang W, Zhang C. Apparent diffusion coefficient values are superior to transrectal ultrasound-guided prostate biopsy for the assessment of prostate cancer aggressiveness. *Acta Radiol* (2017) 58:232–9. doi: 10.1177/0284185116639764

38. Yuan Y, Qin W, Buyyounouski M, Ibragimov B, Hancock S, Han B, et al. Prostate cancer classification with multiparametric MRI transfer learning model. *Med Phys* (2019) 46:756–65. doi: 10.1002/mp.13367

39. Yang X, Liu C, Wang Z, Yang J, Min HL, Wang L, et al. Co-Trained convolutional neural networks for automated detection of prostate cancer in multi-

parametric MRI. *Med Image Anal* (2017) 42:212–27. doi: 10.1016/j.media.2017.08.006

40. Wong T, Schieda N, Sathiadoss P, Haroon M, Abreu-Gomez J, Ukwatta E. Fully automated detection of prostate transition zone tumors on T2-weighted and apparent diffusion coefficient (ADC) map MR images using U-net ensemble. *Med Phys* (2021) 48:6889–900. doi: 10.1002/mp.15181

41. Netzer N, Weißer C, Schelb P, Wang X, Qin X, Görtz M, et al. Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: Effects of cohort size and heterogeneity. *Invest Radiol* (2021) 56:799–808. doi: 10.1097/RLI.0000000000000791

42. Padhani AR, Turkbey B. Detecting prostate cancer with deep learning for MRI: A small step forward. *Radiology* (2019) 293:618–9. doi: 10.1148/radiol.2019192012

43. Yu X, Lou B, Shi B, Winkel D, Szolar D. (2020). False positive reduction using multiscale contextual features for prostate cancer detection in multi-parametric MRI scans, in: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA, USA: IEEE. pp. 1355–9. doi: 10.1109/ISBI45749.2020.9098338

44. Min Z, Bianco FJ, Yang Q, Rodell R, Yan W, Barratt D, et al. (2021). Controlling false Positive/Negative rates for deep-Learning-Based prostate cancer detection on multiparametric MR images, in: *25th UK Conference on Medical Image Understanding and Analysis*. Switzerland: Springer Nature. pp. 56–70. doi: 10.48550/arXiv.2106.02385

45. Liu X, Han C, Cui Y, Xie T, Zhang X, Wang X. Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate cancer based on deep learning. *Front Oncol* (2021) 11:773299. doi: 10.3389/fonc.2021.773299

Frontiers in Oncology

Advances knowledge of carcinogenesis and tumor progression for better treatment and management

The third most-cited oncology journal, which highlights research in carcinogenesis and tumor progression, bridging the gap between basic research and applications to imrpove diagnosis, therapeutics and management strategies.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact