Three-dimensional convolutional neural network model to identify clinically significant prostate cancer in transrectal ultrasound videos: a prospective, multi-institutional, diagnostic study

Yi-Kang Sun,[a,i] Bo-Yang Zhou,[a,i] Yao Miao,[b,c,i] Yi-Lei Shi,[d,i] Shi-Hao Xu,[e] Dao-Ming Wu,[f] Lei Zhang,[d] Guang Xu,[b,c] Ting-Fan Wu,[g] Li-Fan Wang,[a] Hao-Hao Yin,[a] Xin Ye,[a] Dan Lu,[a] Hong Han,[a] Li-Hua Xiang,[b,c,****] Xiao-Xiang Zhu,[h,***] Chong-Ke Zhao,[a,**] and Hui-Xiong Xu,[a,*] China Alliance of Multi-Center Clinical Study for Ultrasound (Ultra-Chance)

[a]Department of Ultrasound, Zhongshan Hospital, Institute of Ultrasound in Medicine and Engineering, Fudan University, Shanghai, China
[b]Department of Medical Ultrasound, Center of Minimally Invasive Treatment for Tumour, Shanghai Tenth People's Hospital, Ultrasound Institute of Research and Education, School of Medicine, Tongji University, Shanghai, China
[c]Shanghai Engineering Research Center of Ultrasound in Diagnosis and Treatment, Shanghai, China
[d]MedAI Technology (Wuxi) Co., Ltd., Wuxi, China
[e]Department of Ultrasonography, The First Affiliated Hospital of Wenzhou Medical University, Zhejiang, China
[f]Department of Ultrasound, Fujian Provincial Hospital, Fujian, China
[g]Bayer Healthcare, Radiology, Shanghai, China
[h]Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany

eClinicalMedicine 2023;60: 102027
Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.102027

Summary

Background Identifying patients with clinically significant prostate cancer (csPCa) before biopsy helps reduce unnecessary biopsies and improve patient prognosis. The diagnostic performance of traditional transrectal ultrasound (TRUS) for csPCa is relatively limited. This study was aimed to develop a high-performance convolutional neural network (CNN) model (P-Net) based on a TRUS video of the entire prostate and investigate its efficacy in identifying csPCa.

Methods Between January 2021 and December 2022, this study prospectively evaluated 832 patients from four centres who underwent prostate biopsy and/or radical prostatectomy. All patients had a standardised TRUS video of the whole prostate. A two-dimensional CNN (2D P-Net) and three-dimensional CNN (3D P-Net) were constructed using the training cohort (559 patients) and tested on the internal validation cohort (140 patients) as well as on the external validation cohort (133 patients). The performance of 2D P-Net and 3D P-Net in predicting csPCa was assessed in terms of the area under the receiver operating characteristic curve (AUC), biopsy rate, and unnecessary biopsy rate, and compared with the TRUS 5-point Likert score system as well as multiparametric magnetic resonance imaging (mp-MRI) prostate imaging reporting and data system (PI-RADS) v2.1. Decision curve analyses (DCAs) were used to determine the net benefits associated with their use. The study is registered at https://www.chictr.org.cn with the unique identifier ChiCTR2200064545.

Findings The diagnostic performance of 3D P-Net (AUC: 0.85–0.89) was superior to TRUS 5-point Likert score system (AUC: 0.71–0.78, $P = 0.003–0.040$), and similar to mp-MRI PI-RADS v2.1 score system interpreted by experienced radiologists (AUC: 0.83–0.86, $P = 0.460–0.732$) and 2D P-Net (AUC: 0.79–0.86, $P = 0.066–0.678$) in the internal and external validation cohorts. The biopsy rate decreased from 40.3% (TRUS 5-point Likert score system) and 47.6% (mp-MRI PI-RADS v2.1 score system) to 35.5% (2D P-Net) and 34.0% (3D P-Net). The unnecessary biopsy rate decreased from 38.1% (TRUS 5-point Likert score system) and 35.2% (mp-MRI PI-RADS v2.1 score system) to 32.0% (2D P-Net) and 25.8% (3D P-Net). 3D P-Net yielded the highest net benefit according to the DCAs.

*Corresponding author. Department of Ultrasound, Zhongshan Hospital, Institute of Ultrasound in Medicine and Engineering, Fudan University, Shanghai, China.
**Corresponding author. Department of Ultrasound, Zhongshan Hospital, Institute of Ultrasound in Medicine and Engineering, Fudan University, Shanghai, China.
***Corresponding author. Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany.
****Corresponding author. Department of Medical Ultrasound, Center of Minimally Invasive Treatment for Tumor, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China.
E-mail addresses: xu.huixiong@zs-hospital.sh.cn (H.-X. Xu), zhaochongke123@163.com (C.-K. Zhao), xiaoxiang.zhu@tum.de (X.-X. Zhu), xianglihua1121@163.com (L.-H. Xiang).
[i]These authors contributed equally to this work.

Interpretation 3D P-Net based on a prostate grayscale TRUS video achieved satisfactory performance in identifying csPCa and potentially reducing unnecessary biopsies. More studies to determine how AI models better integrate into routine practice and randomized controlled trials to show the values of these models in real clinical applications are warranted.

Funding The National Natural Science Foundation of China (Grants 82202174 and 82202153), the Science and Technology Commission of Shanghai Municipality (Grants 18441905500 and 19DZ2251100), Shanghai Municipal Health Commission (Grants 2019LJ21 and SHSLCZDZK03502), Shanghai Science and Technology Innovation Action Plan (21Y11911200), and Fundamental Research Funds for the Central Universities (ZD-11-202151), Scientific Research and Development Fund of Zhongshan Hospital of Fudan University (Grant 2022ZSQD07).

Keywords: Clinically significant prostate cancer; Ultrasound video; Three-dimensional convolutional neural network; Two-dimensional convolutional neural network

Research in context

Evidence before this study

Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer-associated mortality among males worldwide. Early and accurate prediction of clinically significant PCa (csPCa) can help urologists to formulate treatment plans and save patients from unnecessary suffering. Transrectal ultrasound (TRUS) has been proven to be a secondary option for screening patients at risk for PCa when prostate multiparametric magnetic resonance imaging (mp-MRI) is not available, but its effectiveness is still limited by intermediate diagnostic efficacy and poor reproducibility. Artificial intelligence may have the potential to solve the problem of TRUS. However, due to the limitations such as operator dependence of TRUS, it has not achieved satisfactory results.

Added value of this study

In this multicenter study, we developed and validated convolutional neural network (CNN) models (3D P-Net and 2D P-Net) to identify csPCa in standardized prostate TRUS videos before biopsy. These models are specifically optimized for TRUS videos and are capable of analysing information from the entire prostate simultaneously. The proposed strategy outperformed the experienced radiologists and may have the potential to help reduce unnecessary biopsies.

Implications of all the available evidence

Our findings showed that the artificial intelligence-based method was able to identify csPCa in TRUS videos. In clinical practice, these CNN models have the potential to avoid unnecessary biopsies. 3D P-Net provides a possible scheme for diagnosing csPCa especially in prostate mp-MRI limited areas or situations, and may have the potential to be applied to other US-based cancer trials.

Introduction

Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer-associated mortality among males worldwide, representing a great challenge to the healthcare system.[1] Prostate biopsy is the standard of care for diagnosing PCa and is widely used in patients with suspected PCa. Based on pathological findings, PCa is classified into clinically significant PCa (csPCa, Gleason grade grouping [GG] $\geq$ 2 or Gleason score $\geq$ 3 + 4 = 7) and clinically insignificant PCa (cisPCa, GG = 1 or Gleason score $\leq$ 3 + 3 = 6).[2] Unlike other cancers, the prognosis of PCa is closely related to GG. csPCa has a worse prognosis and requires timely intervention and treatment, while cisPCa has a better prognosis (7% mortality at 15-year follow-up), and current guidelines mainly recommend active surveillance.[3] A prostate biopsy can detect csPCa and is widely used in clinical practice; however, it is inevitably associated with increased harm, such as overdiagnosis, complications, and treatment for indolent disease.[4–6] If the investigators cannot accurately determine which patients have a high risk of csPCa and rely primarily on raising biopsy rates to ensure csPCa detection, it will inevitably lead to more unnecessary biopsies (over detection of benign prostatic hypertrophy or cisPCa), increase the pain and burden of patients and cause a great waste of medical resources. Therefore, exploring methods for early detection and accurate identification of patients with csPCa to help optimize prostate biopsy procedures are crucial for csPCa patients' prognosis, surveillance, and management.

Current guidelines endorse the application of prostate multiparametric magnetic resonance imaging (mp-MRI) as the first-line tool for biopsy optimisation.[3] Its irreplaceable value is not only reflected in the diagnosis

of PCa, but also in the localization, staging and prognosis of PCa. However, some factors prevent patients from benefit. One important reason is the low availability and accessibility of prostate mp-MRI in many healthcare systems. Even in developed countries such as the United States, the use of prostate mp-MRI before biopsy has grown slowly over the past five years and remains very low because of poor availability.[7–9] The high price of prostate mp-MRI may be another possible reason. In Europe, for example, where prostate mp-MRI is used more frequently, about 78% of respondents cited high costs as a reason for not using prostate mp-MRI.[10–12] Other factors, such as contraindications to prostate mp-MRI and variability in scan quality or reporting, may also affect its frequency of use.[13–15]

Transrectal ultrasound (TRUS) has been used to examine prostate tissue for more than 40 years. TRUS is another commonly used imaging modality for PCa diagnosis. TRUS has been proven to be a secondary option for screening patients at risk for PCa when prostate mp-MRI is not available.[16] The cost of TRUS devices is approximately one-fifteenth of modern MRI body scanners. Furthermore, TRUS devices are available to be transferred between institutions. Although TRUS has the advantages of lower initial expenditure and convenient and real-time imaging, it is difficult for clinicians to identify and diagnose PCa on TRUS images.[17] The performance of grayscale TRUS varies widely, with sensitivities ranging from 8% to 88% and specificities ranging from 42.5% to 99%.[18] The moderate diagnostic efficacy and poor repeatability of TRUS for diagnosing PCa limit its clinical application. Therefore, to truly benefit PCa patients from TRUS, methods should be refined to improve the performance of TRUS for PCa diagnosis and make the results more stable and predictable.

In recent years, artificial intelligence (AI) techniques have been providing significant improvements in various medical tasks, which can perform similar to human or even better in different domains of application such as tumor detection, classification and prognosis prediction.[19,20] Compared with experts, AI has the potential to reflect not only holistic tumour morphology but also capture task-specific and granular radiological patterns that cannot be detected by the naked eye.[21] Previous studies have shown that this method can help predict PCa based on US prostate imaging.[22,23] Nevertheless, these retrospective studies mainly concentrated on the analysis of handcrafted features on a single US image, had small sample sizes, and lacked external validation to ensure the reliability of their results. Compared with other AI technologies, convolutional neural networks (CNNs) achieve outstanding performance in medical image analysis tasks because they are specifically designed for image recognition tasks. CNN-based image analysis can establish a direct link between complicated medical imaging data and disease prediction.[24,25] In addition, primary PCa is often multifocal.[26] Therefore, the use of a single image for PCa diagnosis may not be sufficient. The analysis of TRUS videos containing the entire prostate may be a viable option. This type of TRUS video can be considered MRI-like volumetric data. However, unlike mp-MRI, the generation of TRUS images is highly operator dependent. Differences in the TRUS images may affect the accuracy of the deep learning model. To minimise the impact of the operator on TRUS image information and improve the repeatability of the model, standardised end-to-end prostate TRUS videos were used for model training and validation in the present study. Studies have shown that mp-MRI data of the entire prostate can be used to predict csPCa and postoperative pathology with the help of 2D and 3D CNNs.[27,28] Theoretically, these approaches may also be applicable in the TRUS. However, only a few studies have focused on this aspect. Whether 2D CNNs or 3D CNNs are more suitable for csPCa diagnosis based on TRUS videos remains open to question.

In this study, we developed and validated frameworks based on 2D and 3D CNNs to analyse standardised end-to-end prostate TRUS videos for identifying csPCa. The framework predicts final pathology results at the patient level without manual annotation. The performance and clinical benefits of these models were further compared with those of radiologists in multicentre datasets.

Methods

Study design

The overall design of this study is illustrated in Fig. 1. In the present multicentre, prospective study, we analysed the clinicopathological characteristics, prostate TRUS imaging data, and prostate mp-MRI data of patients with suspected PCa who underwent prostate biopsy at four institutions. The prospective study was registered at https://www.chictr.org.cn (No. ChiCTR2200064545).

Ethical approval

The institutional ethics committee approved this multicentre study (approval number: SHYS-IEC-5.0/22K212/P01), and informed consent was obtained from all patients. All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Sample size calculating

Sample size required for methodology reliability verification: Sample size is calculated by sensitivity and specificity. The formula for estimating the sample size is as follows:

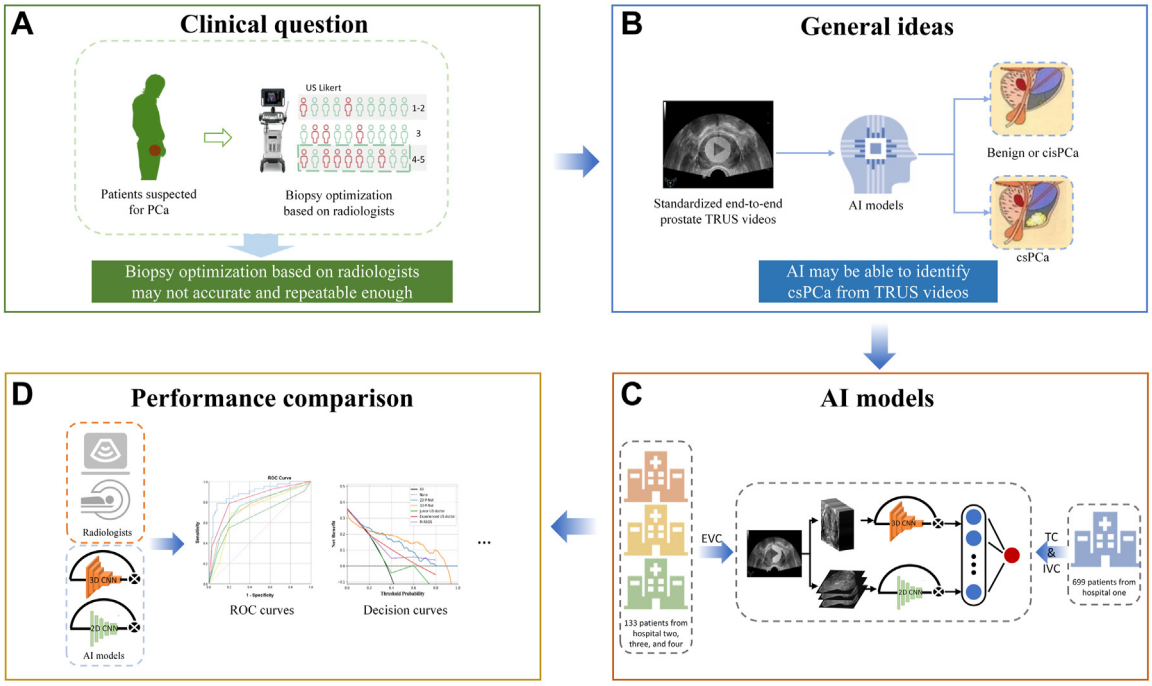

Fig. 1: **The overall design of the study.** (A) Biopsy optimisation by radiologists may not be accurate and repeatable enough. (B) Deep learning-based models for early prediction of csPCa based on standardised end-to-end prostate TRUS videos were constructed. (C) Patients enrolled from Hospital one (Shanghai Tenth People's Hospital) were used as the training cohort, while others recruited from Hospital two (First Affiliated Hospital of Wenzhou Medical University), Hospital three (Fujian Provincial Hospital), and Hospital four (Zhongshan Hospital, Fudan University) were used as external validation cohorts. (D) Predictive performance is compared using AUCs and DCAs. TRUS: transrectal ultrasound; PCa: prostate cancer; csPCa: clinically significant prostate cancer; cisPCa: clinically insignificant prostate cancer, AI: artificial intelligence, TC: training cohort; IVC: internal validation cohort; EVC: external validation cohort; ROC: receiver operating characteristic; AUC: area under the curve; DCAs: decision curve analyses.

$$N = \left(\frac{\mu_\alpha \times \sqrt{p_0 \times (1 - p_0)} + \mu_\beta \times \sqrt{p \times (1 - p)}}{p - p_0} \right)^2$$

N is the required sample size. α is 0.05, and β is 0.2 (the expected test power was 80%). μ_α and μ_β are divided into quantiles of the normal distribution function corresponding to significance level and power. p is the estimated value of the expected sensitivity or specificity, p_0 is the lowest standard of clinically acceptable sensitivity or specificity, and the sample size of the case group or the non-case group is estimated by sensitivity and specificity, respectively.

According to the pre-experimental data, the expected sensitivity is 67.8%. The minimum criterion for clinically acceptable sensitivity is 80%, and the sample size required according to the above formula is 121. Since a certain proportion of training set samples is required for AI research, a total of 605 patients are needed when the training and validation set are allocated according to the 4:1 ratio. Considering the need of the highest dropout rate of about 20%, 726 patients are needed.

Datasets

Consecutive patients who underwent prostate biopsy between January 2021 and December 2022 were prospectively enrolled from the Shanghai Tenth People's Hospital (Hospital one) as training and internal validation cohorts. The external validation cohorts included the First Affiliated Hospital of Wenzhou Medical University (Hospital two), Fujian Provincial Hospital (Hospital three), and Zhongshan Hospital, Fudan University (Hospital four).

The same patient inclusion and exclusion criteria were applied in participated four institutions. The inclusion criteria for the study sample were as follows: (a) underwent TRUS-guided prostate biopsy and (b) accepted TRUS prostate examination prior to biopsy. The exclusion criteria were as follows: (a) history of treatment for PCa (antihormonal therapy, radiation therapy, local therapy, and prostatectomy) before biopsy, (b) presence of other pathological types, and (c) incomplete video data of prostate TRUS examination.

Finally, 819 consecutive patients who underwent TRUS-guided prostate biopsy at hospital one were prospectively enrolled. Among these, 699 patients were included in the study. The other 120 patients were

excluded because of the following reasons: (a) A total of 110 patients were excluded due to incomplete data (including 31 had no TRUS video stored; 48 did not include the whole prostate in the retained TRUS video, 29 had data corruption or loss during storage or transmission process, and 2 had incomplete pathological results due to damage or improper processing of the biopsied tissue), (b) seven patients were excluded due to previous treatment for PCa before biopsy, and (c) three patients were excluded due to the presence of other pathological types. Among them, 559 patients were allocated to the training and test cohorts, and 140 patients were allocated to the internal validation cohort through random selection. Additionally, 138 consecutive patients were prospectively enrolled in the external validation cohort. Four patients were excluded because of incomplete data, and one was excluded because of treatment for PCa before the biopsy. Ultimately, 133 males were included in the external validation cohort (53 patients from Hospital two, 55 from Hospital three, and 25 from Hospital four) (Fig. 2). The process of database establishment is shown in Supplementary Appendix S1.

Prostate TRUS examination protocol and 5-point Likert score system assessment

The TRUS prostate examination protocol used in this study was a transverse-section grayscale TRUS video of a continuous scan of the whole prostate gland from the base to the apex to avoid missing information. This grayscale TRUS video was recorded before each TRUS-guided prostate biopsy and collected in a database. A grayscale TRUS video requires the following: (a) adjust the appropriate depth and gain to keep the prostate clearly seen in the centre of the TRUS images, (b) ensure that the scan is performed at a constant rate, whether suspicious lesions are seen or not, and (c) ensure that the entire prostate is scanned completely. Radiologists with more than eight years of experience in performing prostate TRUS examinations at each hospital independently completed the TRUS examination and TRUS video collection process (S.W. in Hospital one, SH.X. in Hospital two, DM.W. in Hospital three, and CK.Z. in Hospital four, respectively). One radiologist (G.X.) with ten years of experience in prostate TRUS completed the quality control process for the prostate TRUS data. The specific models of the TRUS machines and probes for each centre can be found in Supplementary Table S1.

All prostate TRUS video data were retrospectively and independently reanalysed by two radiologists based on the TRUS 5-point Likert score system (1, normal appearance [homogeneous, echogenic outer gland]; 2, probably normal [minimal heterogeneity of the outer gland]; 3, indeterminate [contour asymmetry or ill-defined echotexture abnormality]; 4, probable carcinoma [focal contour bulge or probable mass]; 5, highly likely carcinoma [focal hypoechoic mass]).[16,29] Based on previous studies, a score equal to or higher than four was determined as the cut-off value for csPCa in this study.[16] One of these two radiologists was a junior radiologist (YK.S., with less than four years of prostate TRUS experience), and the other was an experienced radiologist (LH.X., with more than seven years of prostate TRUS experience). All these radiologists were blinded to the final pathology. They face to face jointly interpreted additional 50 cases to standardize the TRUS 5-point Likert scoring assignment at the beginning of the study.

Mp-MRI PI-RADS v2.1 score system assessment

All examinations were performed using 3.0-T MRI systems with a transabdominal external phased-array coil.

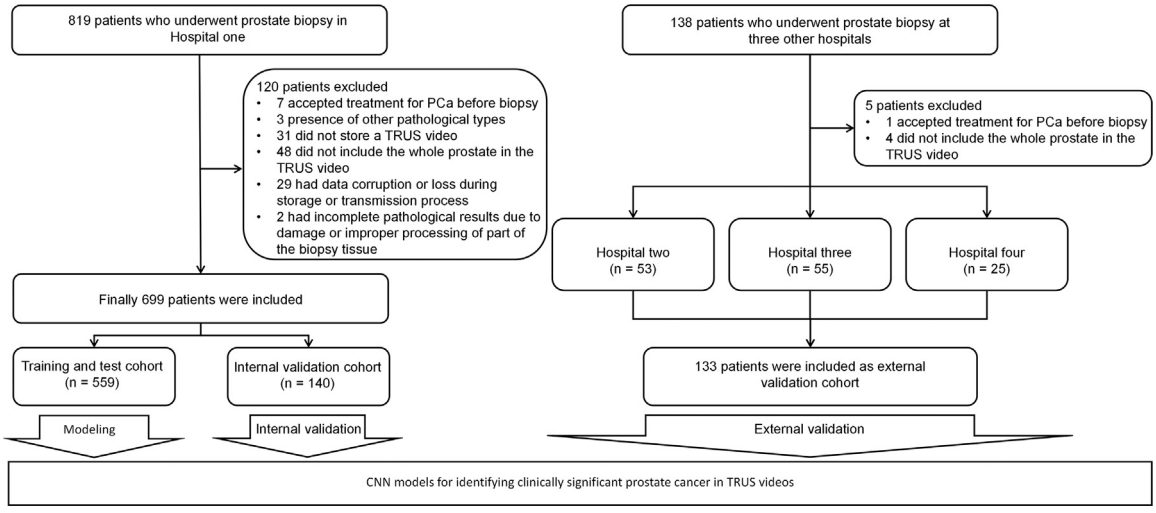

Fig. 2: **Flowchart of the patient enrolment procedure.** TRUS: transrectal ultrasound; PCa: prostate cancer.

Prostate mp-MRI included T1-weighted imaging, T2-weighted imaging, diffusion-weighted imaging, apparent diffusion coefficient, and dynamic contrast enhancement.

The interpretation of prostate mp-MRI PI-RADS v2.1 score system was performed retrospectively by radiologists with more than five years of diagnostic experience in prostate disease in each centre (G.X. and BH.Z. in Hospital one, Q.B. in Hospital two, HR.L. in Hospital three, and XW.Z. in Hospital four).[30] These prostate mp-MRI images were then reanalysed by another junior radiologist (BY.Z., less than 4 years of prostate mp-MRI experience). The likelihood of PCa was assessed according to mp-MRI PI-RADS v2.1 score system on scales from 1 to 5 (1, highly unlikely; 2, unlikely; 3, equivocal; 4, likely; 5, highly likely). Based on previous studies, a score equal to or higher than 4 was determined as the cut-off value for csPCa in this study.[31] When reanalysing the images, all radiologists were blinded to the results of final pathology. With the help of webinar, they jointly interpreted additional 50 cases to standardize the mp-MRI PI-RADS v2.1 scoring assignment at the beginning of the study.

TRUS-guided prostate biopsy and radical prostatectomy

The indications for TRUS-guided prostate biopsy were as follows: (a) abnormally elevated prostate-specific antigen (PSA) level (>4 ng/mL); (b) gradually increasing PSA level by 0.75 ng/mL/year; or (c) positive finding results in digital rectal examination, TRUS examination, or mp-MRI examination. A TRUS-guided prostate biopsy was performed using US equipment equipped with an intracavitary transrectal probe. Each patient underwent TRUS-guided prostate biopsy using the transperineal or transrectal approach, with a core number ranging from 12 to 20, including 12 cores for sextant-specific systematic biopsy and four cores for targeted biopsy in each suspicious region of prostate TRUS and/or mp-MRI. Cognitive or software fusion-targeted prostate biopsies can be applied to suspicious lesions found on prostate mp-MRI.

The indications for radical prostatectomy (RP) were as follows: (a) age $\leq$ 75 years or $\geq$ ten years of remaining life expectancy, (b) general medical condition suitable for surgery, (c) no distant metastases found on preoperative evaluation, and (d) GG $\geq$ 1 (Gleason score $\geq$ 3 + 3 = 6) in TRUS-guided prostate biopsy but not suitable for active surveillance.

Pathological ground truth

Postoperative histological pathology results after biopsy or surgery were regarded as the gold standard. The Gleason scoring system was adopted according to the International Society of Urological Pathology 2005 and 2014 consensus conferences.[32] PCa with GG $\geq$ 2 (Gleason score $\geq$ 3 + 4 = 7) was considered as csPCa in this study. The correspondence between GG and Gleason scores is shown in Supplementary Table S2. Based on the pathology results, the patients were divided into two groups: csPCa and non-csPCa.

Prostate gland segmentation in grayscale TRUS video

The pre-processing procedure of TRUS video data is shown in the Supplementary Appendix S2. To avoid the influence of other tissues around the prostate gland on the results of this study, the prostate gland was segmented in each frame of the TRUS videos before proceeding to the next step. Therefore, we designed a 2D image-based segmentation U network (Efficient-Unet) to help complete the segmentation of the prostate glands (Supplementary Appendix S3). Firstly, the TRUS videos of 186 patients were selected randomly from the 699 patients dataset. Furthermore, the 186 videos were randomly split into the training and validation sets with an 8:2 ratio for developing Efficient-Unet. Based on the size of the prostate gland, 20 to 40 frames (29 frames on average) of images were evenly selected from each TRUS video. These TRUS images were annotated by a radiologist (Y.M.) who had been engaged in prostate TRUS diagnosis for more than three years as the gold standard for Efficient-Unet. Fig. 3 shows detailed information on the development of the Efficient-Unet model architecture. All experiments were conducted using the PyTorch-1.10.1 deep learning framework (https://github.com/pytorch/pytorch/tree/v1.10.1).

2D CNN and 3D CNN models development and validation

Two types of CNN models (2D P-Net and 3D P-Net) were developed for analysing TRUS videos more comprehensively. Commonly used networks (ResNet 50, EfficientNet b0, and DenseNet121 and their corresponding 3D versions) were selected as backbone for 2D P-Net and 3D P-Net (Supplementary Appendices S4–S6). For a fair comparison, a decision algorithm was used to integrate the predictions of all 2D frames to ensure 2D P-Net can yield prediction at patient level. Subsequently, the performance of the 2D P-Net and 3D P-Net were evaluated at patient level. To prevent overfitting of the final models and reduce biased results, stratified 5-*fold* cross-validation was performed (Supplementary Appendix S7) in the training and test cohorts. The final result was the average of models of each fold. CNN models in this study were developed based on the TRUS videos processed by Efficient-Unet (Supplementary Appendix S8).

The TRUS images used for 2D P-Net training and validation in this study were extracted from the TRUS videos. The 2D P-Net comprehensively analyzes each frame of the input image and finally makes a binary judgement of the presence or absence of csPCa at the patient level. The optimal scale of the input

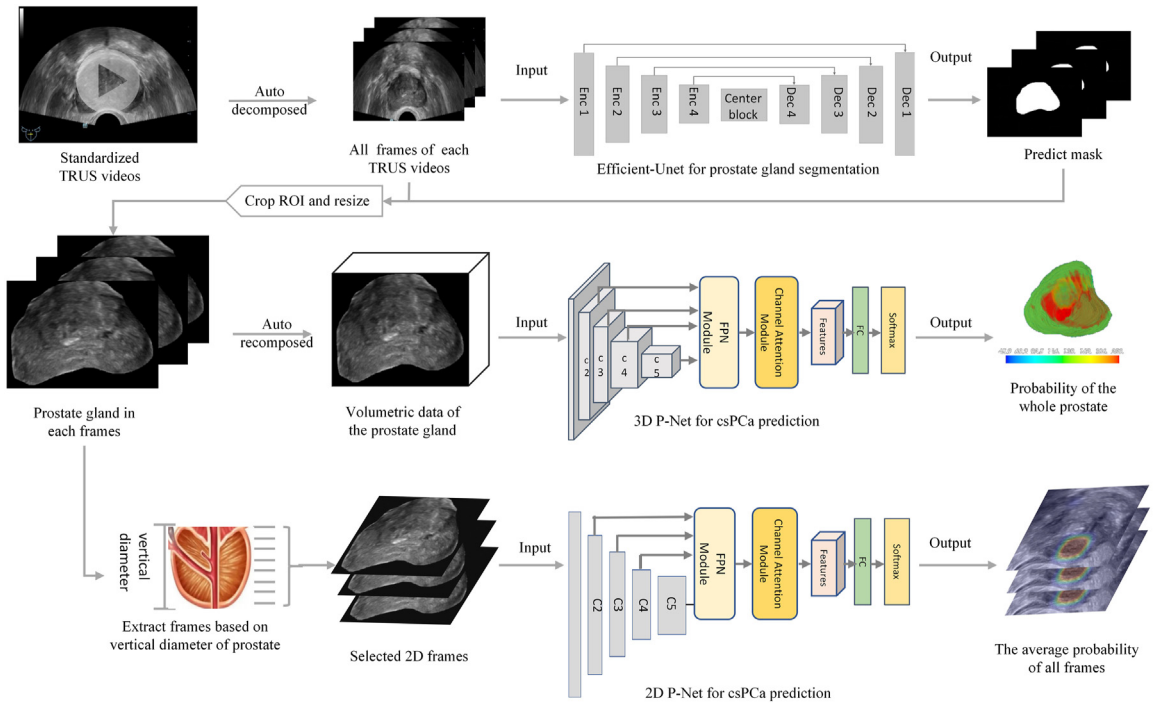

Fig. 3: **The workflow in the development of three CNN models (Efficient-Unet, 2D P-Net, and 3D P-Net) for automated csPCa identification. (TRUS: transrectal ultrasound, csPCa: clinically significant prostate cancer).**

configurations for extracting frames was determined. The TRUS video were sampled at different intervals to obtain 2D frames. Each patient's TRUS video contained a complete prostate scan. The intervals ranged from 0 mm, 3 mm, 6 mm, 9 mm, and 12 mm. The interval of 0 mm indicated that all video frames were used.

Standardised TRUS videos of the entire prostate were used for the training and validation of the 3D P-Net. 3D P-Net can jointly learn temporal and spatial features in the entire TRUS video, which is theoretically suitable for TRUS video analysis. Based on the training of TRUS videos labelled with pathological results, 3D P-Net can eventually make a binary classification judgement of csPCa at the patient level based on each TRUS video.

To reduce the impact of data volume on the model effect, in addition to the traditional data extension methods (random horizontal flip, random vertical flip, and random erase), the self-supervised learning (SSL) methods have been incorporated into the training process of these models (Supplementary Appendix S9). This may avoid excessive attention to material, texture, and other useless features and has been proven to reduce the training cost of CNNs and improve the effectiveness of CNN models.[33] In addition, to make the model more suitable for video data and more effective, a feature pyramid network (FPN) and squeeze-and-excitation networks (SEN) were introduced and applied to the running process of CNNs (Supplementary Appendices S10 and S11). FPN can simultaneously utilise the high resolution of low-level features and the high semantic information of high-level features and can help overcome the varying sizes of prostate glands in TRUS videos. SEN can adaptively adjust the feature weight of each frame and model the internal dependency between video frames, helping the models to focus better on suspicious areas in TRUS videos.

With the help of these networks, CNN models can distinguish the key features from the input information more comprehensively and accurately and further strengthen these key features to improve the performance of the models. The workflow for the development of the three CNN models (Efficient-Unet, 2D P-Net, and 3D P-Net) is shown in Fig. 3. All experiments were conducted using the PyTorch-1.10.1 deep learning framework (https://github.com/pytorch/pytorch/tree/v1.10.1). The cut-off values of these models were based on the mean value of Youden index in the ROC curve of the 5-*fold* cross-validation results of the training cohort.

Clinical parameters combined 2D P-Net and 3D P-Net models development and validation

To investigate the impact of clinical data in improving diagnostic performance, the clinical data were merged to 2D P-Net and 3D P-Net models, respectively. The

univariate and multivariable analysis was used to determine independent predictors of clinical parameters including total PSA, free PSA, PAS density, family history, and previous negative biopsies in identifying csPCa. Then, a logistic regression classifier was trained based on the key clinical parameters and the output probabilities of imaging predictors based on 2D P-Net and 3D P-Net models to obtain the clinical nomogram. The cut-off value of this logistic regression classifier was based on the Youden index in the ROC curve of the training cohort.

Heat map generation

To better interpret the CNNs prediction results, the gradient-weighted class activation mapping (Grad-CAM) method was used to generate a heat map.[34] Heat maps can visualise the most indicative areas of each frame of the TRUS videos or images to interpret the predictive mechanism of the CNN model, which reflects the contribution of each pixel in these images to the prediction of csPCa. All heat maps were produced by applying the packages OpenCV-python-4.6.0 (https://github.com/opencv/opencv/tree/4.6.0) and Mayavi-4.7.3 (https://github.com/enthought/mayavi/tree/4.7.3).

Statistical analysis

All statistical analyses were performed using SPSS (version 22.0, IBM Corporation, Armonk, USA) and Python (version 3.6.13, Python Software Foundation, State of Delaware, USA). The graphs and charts were created based on Matplotlib 3.3.2 (https://github.com/matplotlib/matplotlib/tree/v3.3.2). There is no allowance for multiplicity in our statistical analyses. Shapiro-Welk test was used to evaluate the normal distribution. Differences in clinical factors were analysed using the chi-square test or *t*-test. The biopsy rate was defined as the percentage of patients who required biopsy among all patients. The unnecessary biopsy rate was defined as the percentage of the non-csPCa among the currently total biopsy-required patients. The intersection over union (IoU) and Dice coefficient were used to evaluate the accuracy of the segmentation. The performance of TRUS 5-point Likert score system, mp-MRI PI-RADS v2.1 score system, 2D P-Net, and 3D P-Net were assessed and compared, in term of the AUC, biopsy rate, and unnecessary biopsy rate. The F1-score was also calculated ($F1 = \frac{2Precision \tilde{n} Reca}{Precision + Recal}$). The Delong's test was used to test the differences in AUCs.[35] Inter-group consistencies were compared using the intraclass correlation coefficient (ICC). DCAs were used to determine the clinical net benefit associated with the use of 2D P-Net and 3D P-Net compared with the TRUS 5-point Likert score system and mp-MRI PI-RADS v2.1 score system.[36] $P < 0.05$ was considered statistically significant difference. $P \geq 0.05$ inferred there wasn't enough evidence for a statistical difference.

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript. All authors have full access to all data and approved the final manuscript for submission.

Results

Patient characteristics

The baseline patient characteristics are shown in Table 1. 300 patients with csPCa (GG ≥ 2) and 532 patients with non-csPCa (GG = 1 or no PCa on biopsy) were identified on the pathological evaluation of biopsy reports or RP specimens. A total of 106 cisPCa patients were pathologically proven by prostate biopsy confirmed as GG = 1 in this study. Among them, 46 (43.4%, 46/106) patients were subject to RP surgery. The surgical selection was based on the urologists and patients' decision. 20 (43.5%, 20/46) of them were due to PSA ≥ 10 ng/mL, 8 (17.4%, 8/46) of them were due to suspicious disease progression detected by prostate mp-MRI, and 18 (39.1%, 18/46) of them were due to patient' preference. All patients in this study underwent TRUS-guided systematic biopsy with or without targeted biopsy. A total of 358 patients underwent TRUS-guided systematic biopsy only. In these patients the detection rate of csPCa was 18.7% (67/358). The remaining 341 patients underwent MRI/TRUS fusion or TRUS targeted biopsy. The detection rate of csPCa in these patients were 39.9% (136/341). Of 341 patients, 190 underwent MRI/TRUS-fusion guided targeted biopsy and 151 underwent TRUS targeted biopsy. The detection rate of csPCa in MRI/TRUS-fusion guided targeted biopsy patients were 51.6% (98/190). The detection rate of csPCa in TRUS targeted biopsy patients were 25.2% (38/151).

Segmentation accuracy of Efficient-Unet

The final training set consisted of 4321 images from 149 videos, and the validation set consisted of 1073 images from 37 videos for Efficient-Unet. In the validation set of 1073 TRUS images, Efficient-Unet and manual prostate segmentations had a Dice coefficient of 0.91 and an IoU of 0.85. The distributions of the Dice coefficient and IoU are presented in Supplementary Fig. S1.

Performance of the 2D and 3D CNN models in identifying csPCa

2D and 3D P-Nets development, 2D-ResNet 50 and 3D-ResNet 50 were chosen as the base architectures because their AUCs were slightly better than that of other CNNs in the 5-fold cross-validation results in the training cohort. By adding the FPN module, SEN module, and SSL method to the baseline model, both 2D and 3D P-Nets achieved higher AUCs

Characteristics	Training and test cohort (Hospital one)	Internal validation cohort (Hospital one)	External validation cohort (Hospital two, three, and four)
Number of patients	559	140	133
Age (y, median, IQR)	70 (65, 76)	70 (65, 75)	71 (65, 76)
Total PSA (ng/mL, median, IQR)	10.1 (6.6, 19.7)	10.1 (6.8, 16.9)	11.1 (7.2, 29.2)
Free PSA (ng/mL, median, IQR)	1.5 (1.0, 2.5)	1.6 (1.1, 2.7)	1.8 (1.1, 4.7)
PSAD (ng/mL.cm^3, median, IQR)	0.2 (0.1, 0.5)	0.2 (0.1, 0.3)	0.2 (0.1, 0.3)
Previous negative biopsies			
Present	3 (0.5)	0 (0.0)	0 (0.0)
Absent	556 (99.5)	140 (100.0)	133 (100.0)
Family history			
Present	28 (5.0)	6 (4.3)	4 (3.0)
Absent	531 (95.0)	134 (95.7)	129 (97.0)
Biopsy outcome (n, %)			
Benign	313 (56.0)	84 (60.0)	79 (59.4)
GG 1	83 (14.8)	16 (11.4)	7 (5.3)
GG 2	48 (8.6)	11 (7.9)	11 (8.2)
GG 3	56 (10.0)	17 (12.2)	15 (11.3)
GG 4	19 (3.4)	3 (2.1)	7 (5.3)
GG 5	40 (7.2)	9 (6.4)	14 (10.5)
Treatment (n, %)			
RP	187 (33.5)	43 (30.7)	10 (7.5)
Age ≤ 75 years	159 (85.0)	35 (81.4)	10 (100.0)
≥10 years of remaining life expectancy	28 (15.0)	8 (18.6)	0 (0.0)
Follow up or other treatment for PCa	372 (66.5)	97 (69.3)	123 (92.5)
RP outcome (n, %)			
GG 1	9 (4.8)	5 (11.6)	1 (10.0)
GG 2	60 (32.1)	15 (34.9)	6 (60.0)
GG 3	54 (28.9)	10 (23.3)	2 (20.0)
GG 4	6 (3.2)	1 (2.3)	0 (0.0)
GG 5	58 (31.0)	12 (27.9)	1 (10.0)
Pathological stage			
≤ pT2c	106 (56.7)	25 (58.1)	9 (90.0)
≥ pT3a	81 (43.3)	18 (41.9)	1 (10.0)
Final pathology result (n, %)			
csPCa	203 (36.3)	50 (35.7)	47 (35.3)
Non-csPCa	356 (63.7)	90 (64.3)	86 (64.7)
cisPCa	49 (13.8)	7 (7.8)	7 (8.1)
Benign	307 (86.2)	83 (92.2)	79 (91.9)

PSA: prostate specific antigen; PSAD: prostate specific antigen density; IQR: interquartile range; GG: Gleason grade grouping; RP: radical prostatectomy; csPCa: clinically significant prostate cancer; cisPCa: clinically insignificant prostate cancer.

Table 1: **The basic characteristics of patients.**

(Supplementary Tables S3 and S4). Among different input configurations to the network, the AUCs of extract frames with a scale of 0 mm, 3 mm, 6 mm, 9 mm, and 12 mm are shown in Supplementary Table S5. The AUC of the 3 mm scale was higher than those of the 0 mm and 6–12 mm scales.

Finally, the 2D P-Net in this study was generated by adding FPN module, SEN module, and SSL method to ResNet50, while the 3D P-Net was generated by adding FPN, SE module and SSL method to 3D-ResNet50. The trained 2D and 3D P-Net were then applied to the internal and external validation cohorts.

Using the optimum cut-off value of 0.471, the sensitivity and specificity of 3D P-Net to predict csPCa were 0.63 (95%CI: 0.48, 0.73) and 0.94 (95%CI: 0.89, 0.99) in the internal validation cohort and 0.81 (95% CI: 0.70, 0.93) and 0.78 (95%CI: 0.69, 0.86) in the external validation cohort, respectively. With the optimum cut-off value of 0.396, the sensitivity and specificity of 2D P-Net to predict csPCa were 0.72 (95%CI: 0.57, 0.82) and 0.88 (95%CI: 0.79, 0.94) in the internal validation cohort and 0.64 (95%CI: 0.50, 0.74) and 0.78 (95%CI: 0.68, 0.85) in the external validation cohort, respectively. The statistically significant

difference only appeared in sensitivity for the external validation cohort (P = 0.039) (Supplementary Table S6).

The AUCs of 3D P-Net in predicting csPCa were 0.89 (95%CI: 0.83, 0.95) and 0.85 (95%CI: 0.78, 0.93) in the internal and external validation cohorts, respectively. The AUCs of 2D P-Net for predicting csPCa were 0.86 (95%CI: 0.80, 0.93) and 0.79 (95%CI: 0.71, 0.87) in the internal and external validation cohorts, respectively (Fig. 4, and Supplementary Fig. S2 and Table S6). The Delong's test did not reveal statistical significance between these AUCs.

The performance of the 2D P-Net and 3D P-Net were also assessed at different centres and across different TRUS equipments. The AUCs of P-Nets in different TRUS equipment datasets ranged from 0.83 (95%CI: 0.72, 0.94) to 0.92 (95%CI: 0.84, 0.98) in the internal validation cohort and from 0.74 (95%CI: 0.50, 0.91) to 0.90 (95%CI: 0.80, 0.98) in the external validation cohort. The performance of these networks was similar in both the internal and external validation cohorts (Supplementary Table S8). Furthermore, we analysed the diagnostic performance of 2D P-Net and 3D P-Net in two different PSA level subgroups to evaluate the impact of different prevalence rate on the PPV and NPV. In general, the high PSA level reflects the high prevalence rate. The results showed an increase in PPV among those patients with a high PSA

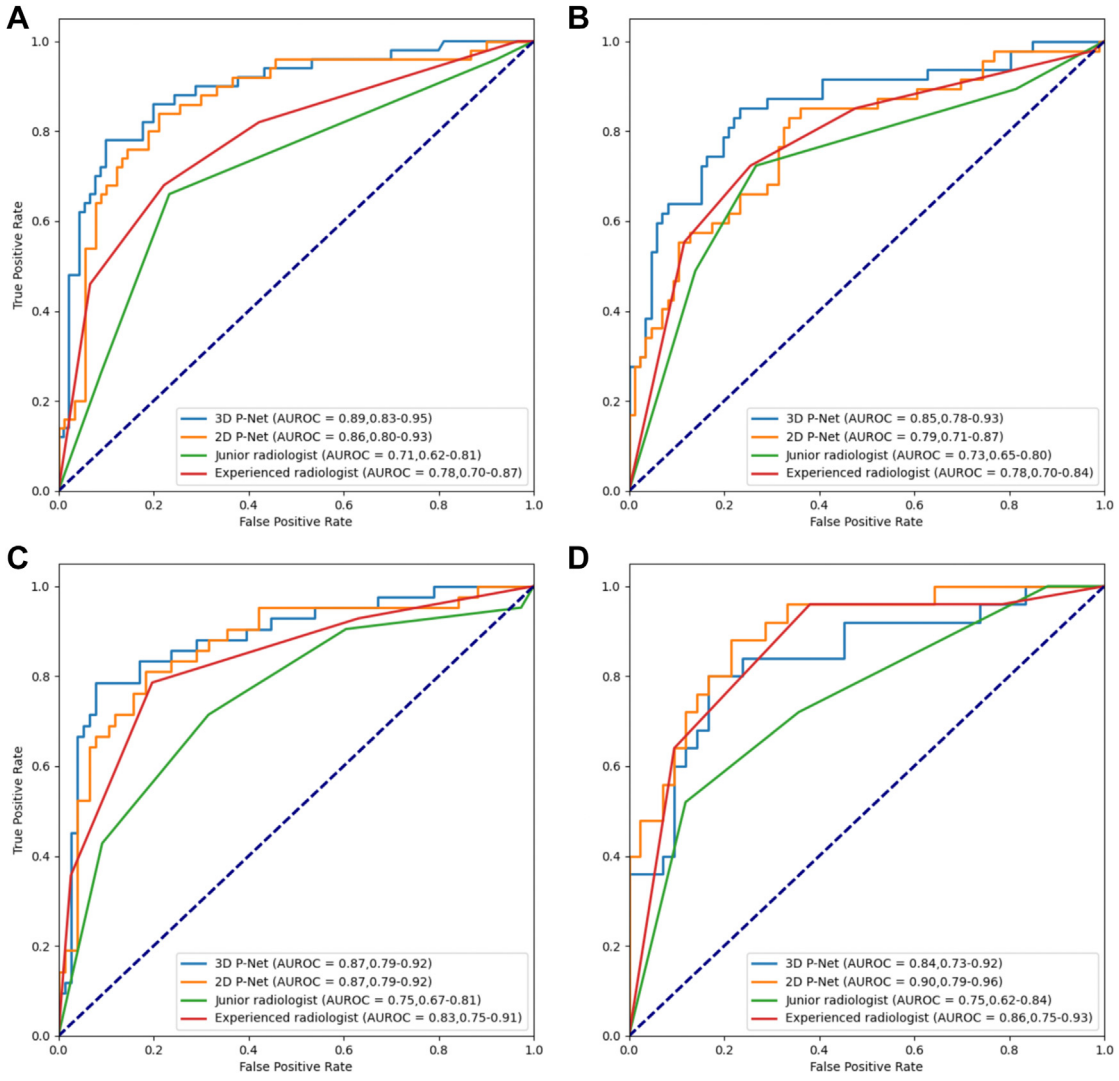

Fig. 4: **Performances for 3D P-Net, 2D P-Net, TRUS 5-point Likert score system, and mp-MRI PI-RADS v2.1 score system.** AUCs of 3D P-Net, 2D P-Net, and TRUS 5-point Likert score system used by junior and experienced radiologists in (A) the internal validation cohort and (B) the external validation cohort. AUCs of 3D P-Net, 2D P-Net, and mp-MRI PI-RADS v2.1 score system in (C) the internal validation cohort and (D) the external validation cohort. TRUS: transrectal ultrasound; mp-MRI: multiparametric magnetic resonance imaging; PI-RADS: prostate imaging reporting and data system; AUROC: area under receiver operating characteristic; AUC: area under the curve.

level and an increase in NPV among those patients with a low PSA level (Supplementary Table S9).

Performance of the TRUS 5-point Likert score system in identifying csPCa

With a cut-off value of four points, the sensitivity and specificity of the junior radiologist using TRUS 5-point Likert score system to predict csPCa were 0.66 (95% CI: 0.54, 0.78) and 0.77 (95%CI: 0.69, 0.84) in the internal validation cohort and 0.72 (95%CI: 0.61, 0.85) and 0.73 (95%CI: 0.64, 0.81) in the external validation cohort, respectively. The sensitivity and specificity of the experienced radiologist using TRUS 5-point Likert score system to predict csPCa were 0.68 (95%CI: 0.56, 0.78) and 0.78 (95%CI: 0.73, 0.86) in the internal validation cohort and 0.72 (95%CI: 0.60, 0.87) and 0.74 (95%CI: 0.66, 0.83) in the external validation cohort, respectively. Only the specificity difference between the radiologists and 3D P-Net for the internal validation cohort was statistically significant ($P < 0.001$ and $P = 0.003$, respectively) (Supplementary Table S6).

For the junior radiologist, the AUCs were 0.71 (95% CI: 0.62, 0.81) and 0.73 (95% CI: 0.65, 0.80) in the internal and external validation cohorts, respectively. For experienced radiologist, the AUCs were 0.78 (95% CI: 0.70, 0.87) and 0.78 (95%CI: 0.70, 0.84) in the internal and external validation cohorts, respectively (Supplementary Table S6). In the internal validation cohort, the AUCs of 2D and 3D P-Nets were better than those of radiologists (all $P < 0.05$) (Fig. 4 and Supplementary Table S6). In the external validation cohort, only the AUC of 3D P-Net was better than that of the radiologists ($P = 0.003$ and $P = 0.039$, respectively), and there was no statistically significant difference between the AUCs of 2D and 3D P-Nets and radiologists ($P = 0.273$ and $P = 0.820$, respectively) (Supplementary Table S6).

The ICC between the experienced radiologist and junior radiologist for interpreting prostate TRUS was 0.46 and 0.69 for the internal and external validation cohorts, respectively.

Performance of the mp-MRI PI-RADS v2.1 score system in identifying csPCa

In the internal validation cohort, 118 patients underwent prostate mp-MRI prior to the biopsy. With a cut-off value of four points, the sensitivity and specificity were 0.71 (95%CI: 0.61, 0.83) and 0.68 (95%CI: 0.57, 0.76) for the junior radiologist and 0.79 (95%CI: 0.67, 0.88) and 0.80 (95%CI: 0.72, 0.88) for the experienced radiologist to predict csPCa, respectively. The specificity difference between radiologists and 3D P-Net was statistically significant ($P < 0.001$ and $P = 0.004$, respectively). The specificity difference between the junior radiologist and 2D P-Net was statistically significant ($P = 0.015$) (Supplementary Table S7). The AUCs were 0.75 (95%CI: 0.67, 0.81) and 0.83 (95%CI: 0.75, 0.91) for the junior and experienced radiologists, respectively. Delong's test showed the AUCs of the 2D and 3D P-Nets were all higher than the junior radiologist ($P = 0.024$ and $= 0.015$, respectively) (Fig. 4 and Supplementary Table S7).

In the external validation cohort, 67 patients underwent prostate mp-MRI prior to the biopsy. With a cut-off value of four points, the sensitivity and specificity were 0.72 (95%CI: 0.56, 0.87) and 0.64 (95%CI: 0.49, 0.80) for the junior radiologist and 0.96 (95%CI: 0.62, 0.92) and 0.62 (95%CI: 0.62, 0.92) for the experienced radiologist to predict csPCa, respectively. Among these data, only the specificity difference between the radiologists and 3D P-Net was statistically significant ($P < 0.001$ and $P = 0.004$, respectively) (Supplementary Table S7). The AUCs were 0.75 (95%CI: 0.62, 0.84) and 0.86 (95%CI: 0.75, 0.93) for the junior and experienced radiologists, respectively. Delong's test showed only the AUC of 2D P-Net was higher than the junior radiologist ($P = 0.020$) (Fig. 4 and Supplementary Table S7).

The ICC between the experienced radiologists and junior radiologist for interpreting prostate mp-MRI was 0.62 and 0.48 for the internal and external validation cohorts, respectively.

Performance of clinical parameters combined 2D and 3D P-Nets in identifying csPCa

After univariate and multivariable analysis, the total PSA and free PSA were identified to be independent predictors of csPCa (Supplementary Table S10). The performance of key clinical parameters combined with 2D and 3D P-Nets are shown in Supplementary Table S11. In the internal validation cohort, a total of 127 patients had complete clinical data. After combining with clinical parameters, the AUC of 2D P-Net changed from 0.87 (95%CI: 0.80, 0.94) to 0.88 (95%CI: 0.82, 0.94). The AUC of 3D P-Net changed from 0.90 (95%CI: 0.84, 0.96) to 0.82 (95%CI: 0.73, 0.89). The differences in these AUCs were not statistically significant ($P = 0.740$ and $P = 0.572$, respectively). In the external validation cohort, a total of 92 patients had complete clinical data. After combining with clinical parameters, the AUC of 2D P-Net changed from 0.82 (95%CI: 0.72, 0.91) to 0.88 (95%CI: 0.81, 0.95). The AUC of 3D P-Net changed from 0.89 (95%CI: 0.80, 0.96) to 0.91 (95%CI: 0.84, 0.97). The differences in these AUCs were not statistically significant ($P = 0.070$ and $P = 0.529$, respectively). The logistic regression classifier nomograms for 2D P-Net + clinical parameters and 3D P-Net + clinical parameters were shown in Supplementary Fig. S4.

Comparison of the required biopsy rate and unnecessary biopsy rate

The ability of 2D P-Net and 3D P-Net in avoiding unnecessary prostate biopsy was also evaluated and compared with that of the TRUS 5-point Likert score system and mp-MRI PI-RADS v2.1 score system. In total, 273 patients were included in the internal and

external validation cohorts. Among these, prostate mp-MRI was performed in 185 patients. The 2D P-Net and 3D P-Net correctly predicted 87.1% (27/31) and 87.1% (27/31) of cisPCa or benign lesions in patients with mp-MRI PI-RADS v2.1 score ≥4, respectively. Meanwhile, 69.0% (29/42) and 73.8% (31/42) of cisPCa or benign lesions in patients with TRUS 5-point Likert score ≥4 were correctly predicted by the 2D P-Net and 3D P-Net, respectively (Supplementary Table S12).

If a prostate biopsy was performed only in patients with TRUS 5-point Likert score ≥4 or mp-MRI PI-RADS v2.1 score ≥4, the proportion of patients required biopsy rate was 40.3% (110/273) and 47.6% (88/185) for the TRUS 5-point Likert score system and the mp-MRI PI-RADS v2.1 score system, respectively, even when image analysis was performed by experts. The TRUS 5-point Likert score system and mp-MRI PI-RADS v2.1 score system would result in 38.1% (42/110) and 35.2% (31/88) of unnecessary biopsies, respectively. However, if only patients predicted to be positive by 2D P-Net or 3D P-Net underwent prostate biopsy, not only did the required biopsy rate decrease to 35.5% (97/273) and 34.0% (93/273) for 2D P-Net and 3D P-Net, respectively, but also the rate of unnecessary biopsies decreased to 32.0% (31/97) for 2D P-Net and 25.8% (24/93) for 3D P-Net, respectively. However, the 2D P-Net missed 17.6% (31/176) patients with csPCa, which was the highest among all these methods (Fig. 5). Based on the result of DCAs, the 3D P-Net had a higher net benefit among all these methods (Supplementary Fig. S3).

Visual interpretation of the 2D P-Net and 3D P-Net

Fig. 6 presents the corresponding heatmaps of the TRUS images or videos of the 2D P-Net and 3D P-Net used to predict csPCa. The different colour distributions reflect the focus of the 2D CNN and 3D CNN models on the most predictive regions and image features in csPCa. The red parts of the heat map indicate that these

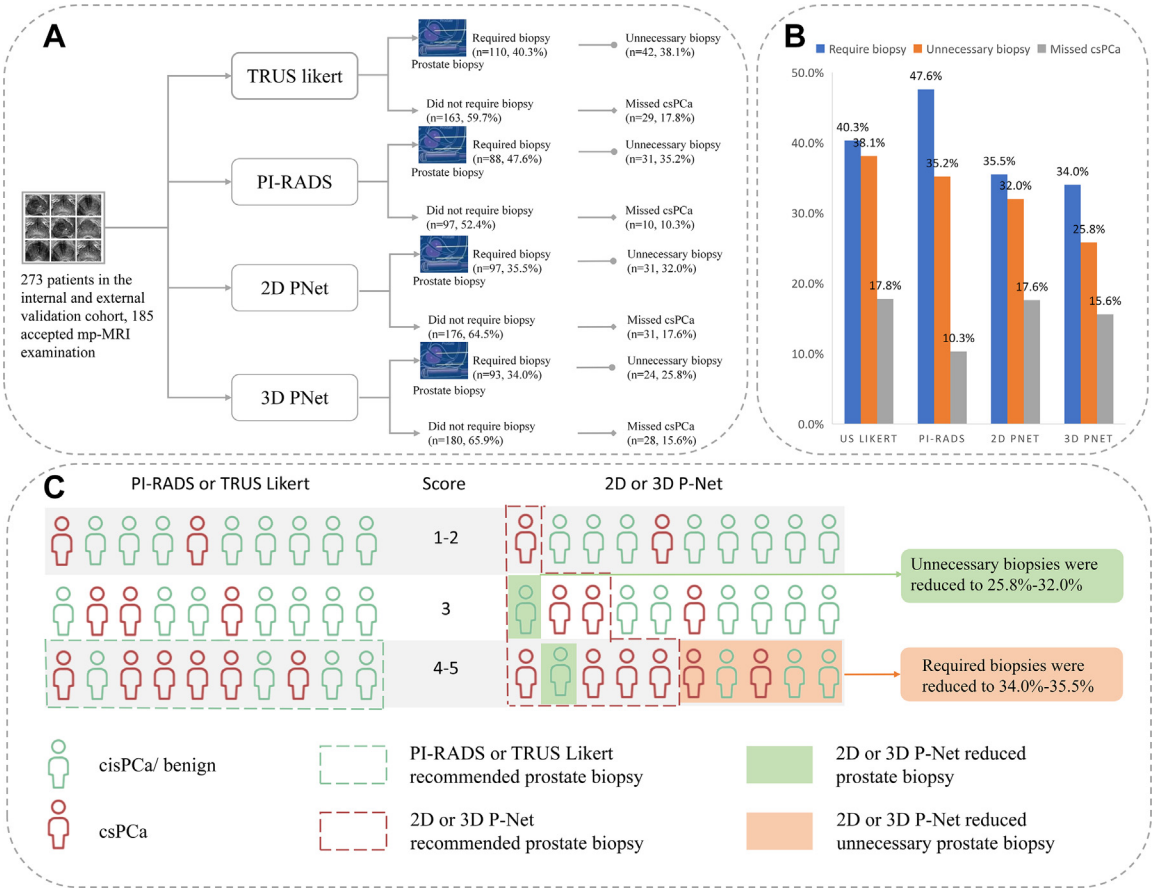

***Fig. 5:* The biopsy rate and unnecessary biopsy rate of 3D P-Net, 2D P-Net, TRUS 5-point Likert score system, and mp-MRI PI-RADS v2.1 score system.** (A, B) The required biopsy, unnecessary biopsy, and missed csPCa rate in the TRUS 5-point Likert score system, mp-MRI PI-RADS v2.1 score system, 2D P-Net, and 3D P-Net, respectively. (C) 2D P-Net and 3D P-Net can reduce the required biopsy rate and unnecessary biopsy rate. TRUS: transrectal ultrasound; mp-MRI: multiparametric magnetic resonance imaging; PI-RADS: prostate imaging reporting and data system.

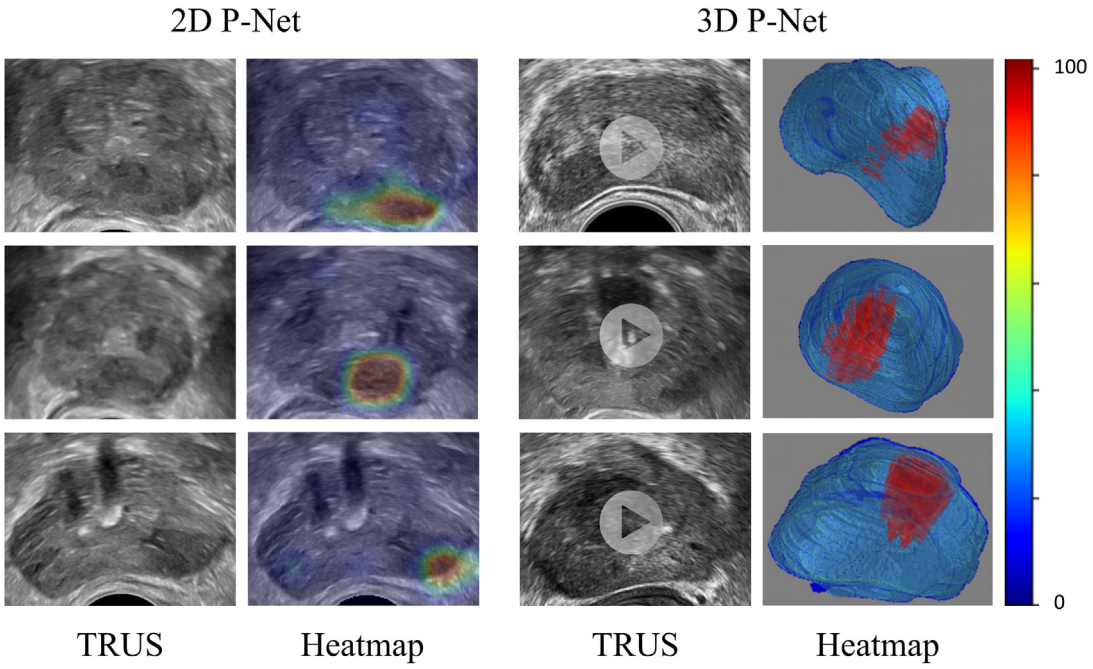

Fig. 6: **The visualisation and interpretation of 2D P-Net and 3D P-Net.** Colour-code heat maps and corresponding transrectal ultrasound (TRUS) images for csPCa patients accurately predicted by 2D P-Net and 3D P-Net.

parts provide more informative features during the network's predictive process. Our results showed that the heatmap of the 2D P-Net and 3D P-Net roughly matched the location of csPCa. The concordance of AI model and prostate mp-MRI in detecting ROI was also compared (Supplementary Appendix S12). The average Dice coefficient between them was 0.54. Considering the possible differences in imaging direction between TRUS and prostate mp-MRI, the consistency of ROI between the two modalities is still acceptable. Therefore, if we could obtain the 2D P-Net or 3D P-Net heatmaps prior to prostate biopsy, we would theoretically be able to identify the location of a suspected csPCa in TRUS videos or images and eventually guide prostate biopsy procedures.

Discussion

In this study, the multiple CNN models (2D P-Net and 3D P-Net) were designed to identify csPCa in prostate TRUS videos. These models were compared with radiologists using the TRUS 5-point Likert score system and mp-MRI PI-RADS v2.1 score system. The results of this study demonstrated that the diagnostic performance of 3D P-Net (AUC: 0.85–0.89) was similar to that of the experienced radiologists using the mp-MRI PI-RADS v2.1 score system (AUC: 0.83–0.86, P = 0.460–0.732) and better than that of the TRUS 5-point Likert score system (AUC: 0.71–0.78, P = 0.003–0.040). The reliability of the models was verified using the heat map results. The performance of the 2D P-Net and 3D P-Net were also assessed at different centres and across different TRUS equipments. The performance of these networks was similar in both the internal and external validation cohorts, indicating good generalizability. Furthermore, our models demonstrated increased PPV for patients with high PSA levels and increased NPV for patients with low PSA levels, suggesting potential benefits in avoiding both overdiagnosis and missed diagnosis among patients with varying prevalence rates.

Prostate mp-MRI is the universally recognized imaging method for csPCa diagnosis. A systematic review showed that the accuracy, sensitivity, and specificity of prostate mp-MRI ranged from 44% to 87%, 58% to 96%, and 23% to 87%, respectively.[37] The finding of this study is consistent with previous reports. However, the high cost of mp-MRI examination and equipment, long learning curves, and moderate inter-reader variability limit the wide application of this technology. The result of this study also showed that the moderate inter-group consistency was achieved between the experienced radiologists and junior radiologist (ICC: 0.48–0.62). Prostate TRUS is a secondary approach for identifying PCa due to the advantages of lower initial expenditure, convenience and so on. Grey et al. applied a 5-point Likert scoring system to TRUS diagnosis and proved that TRUS could be an alternative to prostate mp-MRI as a screening test for patients at risk of PCa.[16] However, the moderate inter-group consistency (ICC: 0.46–0.69) in this study indicates that this system may still need to

be further upgraded to address the problem of relatively low reproducibility.

AI could be a viable solution. As a rapidly advanced computer technology, it has proven to be valuable for medical image analysis, especially in PCa diagnosis.[22,23,27,38–40] However, few studies have focused on identifying csPCa using TRUS data. Wildeboer et al. used radiomics-based methods to analyse handcrafted features on multiparameter US images. They found that the best-performing single parameter yielded AUCs of 0.69 and 0.76, respectively.[23] Nonetheless, their study was based on retrospective data with a small sample size from a single centre, and the features used were handcrafted. The small sample size and high data heterogeneity limit the application and effectiveness of AI in the field of prostate TRUS. Our study was based on standardised end-to-end TRUS videos, which were able to minimise the influence of the operator on the TRUS image information and improve the repeatability of the model. The 2D and 3D P-Nets in this study were based on prospective data with external validation. The results of this study proved that, although relying only on grayscale TRUS, the AUCs of predicting csPCa can reach 0.79–0.89. Our findings suggest that it may be simple and feasible using standardised prostate TRUS videos to better fit AI models.

Establishing AI models based on TRUS videos has always been challenging. Wang et al. used a machine learning-based method to extract features from TRUS video clips to predict csPCa and achieved an AUC of 0.78.[41] However, these approaches required time-consuming manual preparation, which may not be suitable for routine clinical use. Our method requires the least amount of manual pre-processing and may be more advantageous for practical applications. Furthermore, deep learning is a state-of-the-art machine learning approach. This may be a better approach for predicting csPCa. Yet, to our knowledge, no study has used deep learning networks to identify csPCa in prostate TRUS videos. In this study, different types of 2D and 3D CNNs were developed to analyse TRUS videos. The results of this study demonstrated that all these schemes could analyse TRUS videos and that the AUCs obtained were not statistically different. However, our results also reflected the large variation in the performance of 2D P-Net across each validation cohort (AUC: 0.73–0.92), while the performance of 3D P-Net was relatively stable (AUC: 0.84–0.91). A possible reason could be that although both 2D and 3D P-Nets can include all frames of the TRUS video in the analysis, 2D P-Net cannot identify the relationship between each frame, which could lead to very different results between different frames of the same video. 3D P-Net can better capture the temporal and spatial feature information in TRUS videos; therefore, it may be able to avoid the above situations and be more robust than 2D P-Net.

The ultimate goal of the pre-biopsy examinations is to minimise the biopsy and unnecessary biopsy rates based on the lowest possible rate of missed diagnoses. The results of this study suggest that AI models can reduce the rate of unnecessary biopsy compared with the mp-MRI PI-RADS v2.1 score system and TRUS 5-point Likert score systems. However, 2D P-Net may increase the missed diagnosis rate of csPCa. Thus, even if its diagnostic performance is similar to 3D P-Net, it may still fall short regarding the true benefit to patients suspected of PCa. This is also reflected in the DCAs, where 3D P-Net yields the highest net benefit. Although the performance of 2D P-Net was not as good as that of 3D P-Net, 2D P-Net could reduce the minimum requirements of software operation and thus reduce the hardware limitations for software usage. It might be an alternative solution for developing regions with limited conditions, and 3D CNN cannot be applied. Moreover, depending on the effect of different network input configurations on the performance of 2D P-Net (using an input scale of 3 mm, the AUC was higher than that of the other input scales), the minimum requirements for software or hardware of 2D P-Net might be further reduced.

Several studies based on prostate mp-MRI data had applied 3D CNNs in the field of PCa.[28,42,43] Saha et al. designed a 3D CNN model with a complex framework based on the characteristics of prostate mp-MRI to guarantee the effectiveness of the model and achieved an AUC of 0.88.[28] Similar to their method, 3D CNN performed well in processing prostate TRUS videos in our study. Our models were specifically optimised for TRUS videos. Based on the video data characteristics, FPN and SEN were included in the AI models. The results of our study showed that the performance of AI models improved with the addition of these networks. In addition, SSL networks were applied to overcome the huge data requirements of CNNs, which has recently been shown to have the ability to address the appetite for data from CNNs.[33] This study also investigated if this could be further improved with the incorporation of clinical data. The AUC values were slight improved from 0.82–0.90 to 0.82–0.91 but no statistical significance (all $P > 0.05$) was present after combination. The possible reason for this is that clinical parameters may not significantly impact the diagnostic performance of the CNN models in the diagnosis of csPCa.

A reliable prediction model is quite important for radiologists, which assist their decision making and ultimately lead to a better diagnosis of disease. However, radiologists not only want to make predictions, but also want to know how the model makes predictions and whether the network can locate suspicious csPCa lesions. This can be easily achieved by visualising the class activation map in 2D CNNs. However, this approach is not applicable to 3D CNNs. Aldoj et al. used the performance of a 3D CNN to correctly predict lesions and

accuracy during training and testing to provide an impression of the network behaviour.[44] Although this method can reflect the performance of the network to some extent, it cannot aid in the localisation of PCa. Our method uses the Grad-CAM method which can display the regions of concern for both 2D and 3D CNNs. Moreover, the heat map of this experiment proved that the focus areas of the CNNs were consistent with the tumour area and the ROI area of prostate mp-MRI. This may prove that our network has the ability to identify csPCa in TRUS videos. Since this study was based on the model established by TRUS data, it may be possible to perform fusion-targeted biopsies based on the heat map results and further reduce the number of biopsy cores.

The central aim of this study is to develop a simple, feasible and generalizable CNN model to optimize the performance of prostate TRUS in the diagnosis of csPCa. Our result indicated that the 3D P-Net yielded a satisfactory performance at the external validation cohort, and potentially aid clinical decision making through reducing unnecessary biopsy rate. It provides a basis for subsequent studies. AI is not meant to replace a reader, but rather to complement and enhance their abilities. Regarding the standard practice for AI, the best practice depends on the purpose and scenario of using AI. For screening tasks, using a single reader and AI tool can increase efficiency. However, for diagnosis of high-risk patients, accuracy is crucial. Therefore, combining AI tool with two human readers and a third arbitrator can improve diagnostic accuracy. Applying AI in the clinical practice could be considered a low-cost value-added activity if it greatly improves standard care. However, it still cannot yet be proven what is the standard practice of AI method (e.g., complementing a reader, the number of readers needed and the order of those readers, or whether an arbiter is needed, etc.) and where AI method might achieve most benefit on the clinical pathway (e.g., during screening, before biopsy, during active surveillance, etc.). Further randomized controlled trials for the clinical application of AI are needed to adjust and optimize the use of AI in these aspects.

Despite these strong results, our study has several limitations. First, although this is a multi-centre study and the selected four hospitals are tertiary referral centres in the southeast of China, the overall dataset lacks ethnic and geographic diversity. Further nationally cross-sectional studies including a sufficient number of representative hospitals in China are needed to confirm the generalisability of our models. Second, various techniques guided biopsy influence the detection rate for csPCa (51.6% in MRI/TRUS fusion targeted biopsy, 25.2% in TRUS targeted biopsy, and 18.7% in TRUS-guided systematic biopsy, respectively). It results in possible bias for performance of P-Nets in this study. Thus, further studies are needed to assess the possible impact of different biopsy methods on the study results. Third, to ensure the accuracy of our models, complete case analysis was used in our study. The patients with incomplete data were excluded, which might result in a selection bias. In our study, these incomplete data were missing at random. Meanwhile, there was no statistically significant difference between the missing cases and enrolled dataset in baseline features (Supplementary Table S13). Fourth, we did not label all lesions in the TRUS video based on pathological results. This could further increase the diagnostic efficacy of the proposed model. However, each TRUS video consists of more than 200 frames and has a high operator dependence on TRUS examinations. It is difficult to accurately label all suspicious prostate lesions based on the pathological results. Instead, imprecise labelling may affect model performance. Therefore, we chose this method, which does not require a tedious manual preparation process, and our results demonstrate that this method is feasible. Finally, the 2D P-Net and 3D P-Net were developed based on grayscale TRUS videos in this study. Although the sample size was calculated, it is still be relatively small for a deep-learning network. Further studies based on large sample and multiparameter TRUS may further improve the performance and reliability of our AI models.

In conclusion, to explore a simple, feasible and generalizable protocol using AI to optimize the performance of TRUS in the diagnosis of csPCa, we developed the 3D P-Net capable of identifying csPCa in TRUS videos and compared their performance with other methods. After multicentre validation, our method yielded a satisfactory performance and offered a meaningful reduction in the unnecessary biopsy recommendation. The system provides a promising scheme based on TRUS for diagnosing csPCa especially in medical institutions where qualified mp-MRI equipment is not available, and might have the potential to be applied to other US-based cancer trials. More studies to determine how AI models better integrate into routine practice and randomized controlled trials to show the values of these models in real clinical applications are warranted.

Contributors

Conceptualization: YL.S., CK.Z., LH.X., HX.X.
Methodology: YK.S., XX.Z., BY.Z., YL.S., TF.W.
Accessing and verifying the underlying data: YK.S., Y.M., L.Z., SH.X., DM.W., G.X., LF.W., HH.Y., X.Y., D.L., H.H.
Investigation: YK.S., CK.Z., BY.Z., YL.S., Y.M., LH.X., HX.X.
Visualization: YK.S., L.Z., YL.S.
Supervision: HX.X., CK.Z., XX.Z., LH.X., YL.S.
Writing—original draft: YK.S., CK.Z., Y.M., L.Z.
Writing—review & editing: CK.Z., HX.X., XX.Z., LH.X., YL.S., BY.Z., TF.W.

Data sharing statement

The data related to patients are not available for public access due to patient privacy requirements but can be obtained from the corresponding author on reasonable request after being approved by the institutional review board.

Declaration of interests
The authors declare that they have no conflict of interest.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (Grant 82202174 and 82202153), the Science and Technology Commission of Shanghai Municipality (Grants 18441905500 and 19DZ2251100), Shanghai Municipal Health Commission (Grants 2019LJ21 and SHSLCZDZK03502), Shanghai Science and Technology Innovation Action Plan (21Y11911200), and Fundamental Research Funds for the Central Universities (ZD-11-202151), Scientific Research and Development Fund of Zhongshan Hospital of Fudan University (Grant 2022ZSQD07).

Appendix ASupplementary data
Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102027.

References
1 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7–33. https://doi.org/10.3322/caac.21708.
2 Matoso A, Epstein JI. Defining clinically significant prostate cancer on the basis of pathological findings. *Histopathology*. 2019;74(1):135–145. https://doi.org/10.1111/his.13712.
3 Mottet N, van den Bergh RCN, Briers E, 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):243–262. https://doi.org/10.1016/j.eururo.2020.09.042.
4 Ilic D, Neuberger MM, Djulbegovic M, Dahm P. Screening for prostate cancer. *Cochrane Database Syst Rev*. 2013;2013(1):Cd004720. https://doi.org/10.1002/14651858.CD004720.pub3.
5 Vickers AJ. Prostate cancer screening: time to question how to optimize the ratio of benefits and harms. *Ann Intern Med*. 2017;167(7):509–510. https://doi.org/10.7326/m17-2012.
6 Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med*. 2009;360(13):1320–1328. https://doi.org/10.1056/NEJMoa0810084.
7 Liu W, Patil D, Howard DH, et al. Impact of prebiopsy magnetic resonance imaging of the prostate on cancer detection and treatment patterns. *Urol Oncol*. 2019;37(3):181.e15–181.e21. https://doi.org/10.1016/j.urolonc.2018.11.004.
8 Kim SP, Karnes RJ, Mwangi R, et al. Contemporary trends in magnetic resonance imaging at the time of prostate biopsy: results from a large private insurance database. *Eur Urol Focus*. 2021;7(1):86–94. https://doi.org/10.1016/j.euf.2019.03.016.
9 Liu W, Patil D, Howard DH, et al. Adoption of prebiopsy magnetic resonance imaging for men undergoing prostate biopsy in the United States. *Urology*. 2018;117:57–63. https://doi.org/10.1016/j.urology.2018.04.007.
10 Couñago F, Sancho G, Gómez-Iturriaga A, Henríquez I. Multiparametric MRI for prostate cancer: a national survey of patterns of practice among radiation oncologists in Spain. *Clin Transl Oncol*. 2018;20(11):1484–1491. https://doi.org/10.1007/s12094-018-1919-z.
11 Saar M, Linxweiler J, Borkowetz A, et al. Current role of multiparametric MRI and MRI targeted biopsies for prostate cancer diagnosis in Germany: a nationwide survey. *Urol Int*. 2020;104(9–10):731–740. https://doi.org/10.1159/000508755.
12 Renard-Penna R, Rouvière O, Puech P, et al. Current practice and access to prostate MR imaging in France. *Diagn Interv Imaging*. 2016;97(11):1125–1129. https://doi.org/10.1016/j.diii.2016.06.010.
13 Tammisetti VS. MR safety considerations for patients undergoing prostate MRI. *Abdom Radiol (NY)*. 2020;45(12):4097–4108. https://doi.org/10.1007/s00261-020-02730-0.
14 Russo RJ, Costa HS, Silva PD, et al. Assessing the risks associated with MRI in patients with a pacemaker or defibrillator. *N Engl J Med*. 2017;376(8):755–764. https://doi.org/10.1056/NEJMoa1603265.
15 Sonn GA, Fan RE, Ghanouni P, et al. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur Urol Focus*. 2019;5(4):592–599. https://doi.org/10.1016/j.euf.2017.11.010.
16 Grey ADR, Scott R, Shah B, et al. Multiparametric ultrasound versus multiparametric MRI to diagnose prostate cancer (CADMUS): a prospective, multicentre, paired-cohort, confirmatory study. *Lancet Oncol*. 2022;23(3):428–438. https://doi.org/10.1016/s1470-2045(22)00016-x.
17 Correas JM, Halpern EJ, Barr RG, et al. Advanced ultrasound in the diagnosis of prostate cancer. *World J Urol*. 2021;39(3):661–676. https://doi.org/10.1007/s00345-020-03193-0.
18 Postema A, Mischi M, de la Rosette J, Wijkstra H. Multiparametric ultrasound in the detection of prostate cancer: a systematic review. *World J Urol*. 2015;33(11):1651–1659. https://doi.org/10.1007/s00345-015-1523-6.
19 Liu Z, Li Z, Qu J, et al. Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res*. 2019;25(12):3538–3547. https://doi.org/10.1158/1078-0432.Ccr-18-3190.
20 Zhou BY, Wang LF, Yin HH, et al. Decoding the molecular subtypes of breast cancer seen on multimodal ultrasound images using an assembled convolutional neural network model: a prospective and multicentre study. *EBioMedicine*. 2021;74:103684. https://doi.org/10.1016/j.ebiom.2021.103684.
21 Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563–577. https://doi.org/10.1148/radiol.2015151169.
22 Liang L, Zhi X, Sun Y, et al. A nomogram based on a multiparametric ultrasound radiomics model for discrimination between malignant and benign prostate lesions. *Front Oncol*. 2021;11:610785. https://doi.org/10.3389/fonc.2021.610785.
23 Wildeboer RR, Mannaerts CK, van Sloun RJG, et al. Automated multiparametric localization of prostate cancer based on B-mode, shear-wave elastography, and contrast-enhanced ultrasound radiomics. *Eur Radiol*. 2020;30(2):806–815. https://doi.org/10.1007/s00330-019-06436-w.
24 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vision*. 2015;115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y.
25 Xu Y, Hosny A, Zeleznik R, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res*. 2019;25(11):3266–3275. https://doi.org/10.1158/1078-0432.Ccr-18-2495.
26 Haffner MC, Zwart W, Roudier MP, et al. Genomic and phenotypic heterogeneity in prostate cancer. *Nat Rev Urol*. 2021;18(2):79–92. https://doi.org/10.1038/s41585-020-00400-w.
27 Shao L, Yan Y, Liu Z, et al. Radiologist-like artificial intelligence for grade group prediction of radical prostatectomy for reducing upgrading and downgrading from biopsy. *Theranostics*. 2020;10(22):10200–10212. https://doi.org/10.7150/thno.48706.
28 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. https://doi.org/10.1016/j.media.2021.102155.
29 Halpern EJ, Ramey JR, Strup SE, Frauscher F, McCue P, Gomella LG. Detection of prostate carcinoma with contrast-enhanced sonography using intermittent harmonic imaging. *Cancer*. 2005;104(11):2373–2383. https://doi.org/10.1002/cncr.21440.
30 Turkbey B, Rosenkrantz AB, Haider MA, 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(3):340–351. https://doi.org/10.1016/j.eururo.2019.02.033.
31 Mazzone E, Stabile A, Pellegrino F, et al. Positive predictive value of prostate imaging reporting and data system version 2 for the detection of clinically significant prostate cancer: a systematic review and meta-analysis. *Eur Urol Oncol*. 2021;4(5):697–713. https://doi.org/10.1016/j.euo.2020.12.004.
32 Humphrey PA. Histopathology of prostate cancer. *Cold Spring Harb Perspect Med*. 2017;7(10):a030411. https://doi.org/10.1101/cshperspect.a030411.
33 Chen Z, Agarwal D, Aggarwal K, et al. Masked image modeling advances 3D medical image analysis. *arXiv*. 2022. https://doi.org/10.1109/WACV56688.2023.00201.
34 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vision*. 2020;128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7.

35 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.

36 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. https://doi.org/10.1177/0272989x06295361.

37 Fütterer JJ, Briganti A, De Visschere P, et al. Can clinically significant prostate cancer be detected with multiparametric magnetic resonance imaging? A systematic review of the literature. *Eur Urol*. 2015;68(6):1045–1053. https://doi.org/10.1016/j.eururo.2015.01.013.

38 Hiremath A, Shiradkar R, Fu P, et al. An integrated nomogram combining deep learning, Prostate Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. *Lancet Digit Health*. 2021;3(7):e445–e454. https://doi.org/10.1016/s2589-7500(21)00082-0.

39 Schelb P, Kohl S, Radtke JP, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. *Radiology*. 2019;293(3):607–617. https://doi.org/10.1148/radiol.2019190938.

40 Bhattacharya I, Seetharaman A, Kunder C, et al. Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework. *Med Image Anal*. 2022;75:102288. https://doi.org/10.1016/j.media.2021.102288.

41 Wang K, Chen P, Feng B, et al. Machine learning prediction of prostate cancer from transrectal ultrasound video clips. *Front Oncol*. 2022;12:948662. https://doi.org/10.3389/fonc.2022.948662.

42 Yang X, Liu C, Wang Z, et al. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. *Med Image Anal*. 2017;42:212–227. https://doi.org/10.1016/j.media.2017.08.006.

43 Le MH, Chen J, Wang L, et al. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. *Phys Med Biol*. 2017;62(16):6497–6514. https://doi.org/10.1088/1361-6560/aa7731.

44 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(2):1243–1253. https://doi.org/10.1007/s00330-019-06417-z.