Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A generalizable deep learning model for preoperative ECE detection in prostate cancer: a multicenter study

Baolin Zhang
The First Affiliated Hospital of Soochow University
Zhenkun Zhang
The First Affiliated Hospital of Soochow University
Danyu Wu
The Third People's Hospital of Xiangcheng District
Haoyang Zhang
The Fourth Affiliated Hospital of Soochow University
Shanfu Lu
Jiaxing Yaohe Technology Co., Ltd
Lifeng Zhang
Changzhou No.2 People's Hospital
Yuanyuan Mi
Affiliated Hospital of Jiangnan University
Zhixin Ling
lzx9860849540126.com

The First Affiliated Hospital of Soochow University

Research Article

Keywords: Extracapsular extension, Multiparametric prostate MRI, Multi-modal 3D attention network, Automated prostate segmentation, Cross-center validation

Posted Date: April 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9292839/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background Accurate preoperative assessment of extracapsular extension (ECE) in prostate cancer is crucial for surgical planning but is limited by inter-reader variability and the poor generalizability of existing models. This study aimed to develop and validate a generalizable deep learning framework based on multiparametric MRI (mpMRI) for automated, cross-center preoperative ECE prediction.

Methods

This retrospective study included 1044 prostate cancer patients from four centers. We proposed an end-to-end workflow comprising automated prostate segmentation (nnU-Net2) and ECE prediction using a novel Multi-Modal 3D Attention Network (M3A-Net). M3A-Net utilized parallel SEResNet101-based encoders with intra-modal attention and contrastive learning to extract features from multi-modal MRI (T2WI, ADC, DWI). Model discrimination (AUC), calibration (Brier score), and clinical utility (decision curve analysis) were evaluated against five baseline deep learning models under a unified protocol, while inter-cohort differences were analyzed via Mann-Whitney U and chi-squared tests.

Results

M3A-Net achieved superior discrimination on the internal test set with the highest AUC (0.873 vs. 0.740 for the second-best) and the lowest Brier score (0.151) against five baseline models. It also demonstrated promising generalization on the external validation set (AUC = 0.850), achieving a clinically relevant balance between high sensitivity (0.867) and moderate specificity (0.683). Decision curve analysis suggested that M3A-Net may offer a greater net clinical benefit compared to all five baseline models.

Conclusion

The proposed M3A-Net framework shows promise for high-precision preoperative ECE prediction with favorable generalization across multi-center data. This model warrants further investigation as a potential tool to assist in surgical planning, thereby potentially improving oncological and functional outcomes.

Introduction

Prostate cancer is a leading malignancy among men, and multiparametric magnetic resonance imaging (mpMRI) has become essential for identifying clinically significant prostate cancer. The Prostate Imaging Reporting and Data System (PI-RADS) v2.1 standardizes image interpretation, improving diagnostic accuracy [1]. Extracapsular extension (ECE) is a critical prognostic factor, as it increases the likelihood of positive surgical margins (PSM) during radical prostatectomy. PSM is linked to higher risks of early biochemical recurrence and local tumor recurrence [2]. Patients with PSM often need adjuvant therapies, such as androgen deprivation therapy (ADT) or radiotherapy. These treatments carry significant side effects, including osteoporosis, cardiovascular issues, urinary incontinence, bowel

dysfunction, and psychological distress [3–6]. When ECE is absent, nerve-sparing radical prostatectomy can preserve neurovascular bundles, supporting sexual function and early recovery of urinary continence [7, 8]. However, when ECE is present, extensive resection of periprostatic tissue is necessary to reduce PSM rates, lower long-term tumor recurrence, and improve survival [9].

In 2019, the American National Cancer Institute (NCI) developed a grading system for assessing ECE on mpMRI. This system offers a standardized, reproducible framework for predicting ECE, similar to PI-RADS [10]. Yet, its accuracy depends on radiologists' experience, MRI scanner performance, and image quality. Significant variability exists among radiologists in diagnosing ECE [11–13]. Additionally, anatomical differences in the prostate gland and tumor regions across patients create challenges for consistent and accurate MRI-based ECE diagnosis. These issues lead to subjectivity and variability, leaving early prostate cancer diagnosis and preoperative ECE assessment as unmet clinical needs.

The inherent limitations of human-centered diagnostic approaches underscore the potential of artificial intelligence (AI) to address these challenges. AI's core strength lies in its ability to consistently and objectively identify complex patterns across large datasets, offering a promising solution to mitigate inter-observer variability and improve diagnostic reliability. By leveraging AI, the field aims not only for marginal improvements but also for addressing the fundamental challenge of diagnostic consistency, which directly impacts the reliability of treatment decisions and, ultimately, patient outcomes.

In recent years, several research teams have developed AI-based models for diagnosing ECE in prostate cancer, showing promising results as potential diagnostic tools [14–20]. For instance, Pegah Khosravi and colleagues reported on the AutoRadAI model, which used data from 1,001 prostate cancer patients for development and validation. This model has the largest sample size to date for predicting ECE. However, it relies solely on T2-weighted axial MRI images and data from a single clinical practice, which may introduce bias. Without external validation, future work should include broader datasets and additional MRI sequences to improve its transferability [14]. Another AI model, Unfold AI, approved by the Food and Drug Administration (FDA) for clinical trial in prostate cancer, demonstrated high diagnostic performance with an AUC of 0.81, sensitivity of 68%, specificity of 76%, balanced accuracy of 72%, positive predictive value (PPV) of 73%, and negative predictive value (NPV) of 70%. However, its results are based on a study of only 147 prostate cancer patients, a limited sample size. The lack of external clinical data validation also hinders its generalizability [15]. Other studies on AI-based models for diagnosing ECE in prostate cancer are limited by small sample sizes, particularly insufficient data from ECE patients [16–18], surgeries done by single surgeon [17, 18] or lack of external center validation [16, 18], increased radiologist workload from manual lesion segmentation [17, 18], or reliance on a single MRI scanner type [19], which reduces result applicability.

A total of 1,044 patients with prostate cancer from four medical centers were enrolled in this study. Notably, 407 of these patients had ECE. Radical prostatectomies were performed by 10 different surgeons. The study utilized imaging data from six different MRI scanners. We developed a fully automated segmentation model that processes multi-modal imaging data to predict the likelihood of

prostate cancer ECE. This large-scale clinical dataset, coupled with the diversity of MRI instruments and surgeons, aims to improve the model's generalizability. Furthermore, the fully automated segmentation model reduces the workload for radiologists. This approach holds promise for providing significant clinical value in the preoperative assessment of ECE.

Materials and methods

Patients' enrollment

Patients who underwent radical prostatectomy (RP) at four medical centers from January 2020 to July 2025 were preliminarily identified. Patients were excluded if they met the following exclusion criteria: 1) the unavailability of 3.0 T MRI images, poor image quality, or mpMRI performed more than 4 weeks preoperatively; 2) history of neoadjuvant therapy (e.g., endocrine therapy, chemotherapy); 3) incidental prostate cancer diagnosed by transurethral resection of the prostate (TURP); 4) prostate biopsy performed at an outside hospital with a lack of detailed pathological reports in our system; 5) final pathology lacking detailed descriptions of extracapsular extension; or 6) prostatectomy performed for benign prostatic hyperplasia (BPH).

Based on the criteria above, a total of 1044 patients were eligible for the study. The breakdown of the finally included cohort was as follows: 639 patients from the First Affiliated Hospital of Soochow University (SUH 1st), 225 from the Second People's Hospital of Changzhou (CZH 2nd), 64 from The Fourth Affiliated Hospital of Soochow University (SUH 4th), and 116 from the Affiliated Hospital of Jiangnan University (JUH). The study flow diagram is shown in Fig. 1.

MRI acquisition, PI-RADS assessment, and Histopathology

All mpMRI exams were performed using 3.0-T MRI scanners equipped with pelvic phased-array coils. The specific scanner models included the MAGNETOM Skyra (Siemens Healthineers), MAGNETOM Verio (Siemens Healthineers), MAGNETOM Vida (Siemens Healthineers), Ingenia 3.0T (Philips Healthcare), Signa Premier (GE Healthcare), and Discovery MR750 (GE Healthcare). The mpMRI protocol included T2-weighted imaging (T2WI) in three planes, axial Diffusion-Weighted Imaging (DWI) and Apparent Diffusion Coefficient (ADC) maps, and Dynamic Contrast-Enhanced (DCE) imaging.

The PI-RADS scoring at each center was conducted by a junior radiologist (> 5 years of experience) and subsequently confirmed by a senior radiologist (> 20 years of experience according to PI-RADS version 2.1. Similarly, all pathological results from both biopsies and radical prostatectomy specimens were first interpreted by a pathologist (> 5 years' experience) and then confirmed by a senior pathologist (> 20 years' experience). The postoperative pathology report was required to detail the ISUP Grade Group [21], seminal vesicle invasion (SVI), extracapsular extension (EPE), perineural invasion (PNI), and lymph node involvement (LNI).

Proposed AI Framework We proposed an end-to-end workflow comprising two stages:

Automated Segmentation

An nnU-Net2 model was trained on a hybrid dataset, integrating 176 private ROI samples with 78 public samples from the QIN-Prostate-Repeatability and Medical Segmentation Decathlon datasets. To ensure the reliability of the automated pipeline, a rigorous selection process was implemented where only cases achieving a Dice Similarity Coefficient (DSC) greater than 0.85 during cross-validation were utilized for the final model construction. This strategy ensured that the downstream ECE prediction was based on highly accurate prostate boundaries. The detailed process was indicated in Fig. 2.

M3A-Net Architecture

The segmented volumes were processed by our proposed M3A-Net. To fully extract unique feature information from each modality, the network employs independent, parallel 3D residual encoders based on the SEResNet101 architecture for T2WI, ADC, and DWI. Each encoder consists of four stages of stacked 3D residual blocks containing 3, 4, 23, and 3 units respectively, progressively expanding the feature dimensions to a high-level 2048-dimensional vector. The architecture incorporates an intra-modal Squeeze-and-Excitation (SE) attention mechanism, allowing the model to adaptively focus on the most critical features for ECE determination through channel-wise reweighting. Finally, these features are concatenated into a unified 6144-dimensional vector for multi-modal fusion and classification. The complete methodological pipeline of our study is illustrated in Fig. 3. (An additional movie file shows this in more detail [see Additional file 1].)

Experimental Setup

Data Splits and Evaluation Framework

A multi-phase validation framework was employed, dividing the dataset into a training set (n = 800), an internal test set (n = 64), and an external validation set (n = 180) from independent medical centers. Model performance was comprehensively assessed in terms of discrimination, calibration, and clinical utility.

Compared Methods

The proposed M3A-Net was benchmarked against five representative baseline models—3D-Swin[22], PAGNet[19], ExCapNet[14], a generic 3D CNN[23], and a ResNet-18 variant[24]—under a unified training and inference protocol using T2WI, ADC, and DWI inputs.

Implementation Details

All experiments were implemented in PyTorch using the MONAI framework. For image preprocessing, multimodal independent normalization was performed using Z-score standardization on non-background regions (signal values > 0), with values clipped to [-3, 3] to minimize the impact of extreme outliers. All input volumes were resampled to a unified spatial resolution of 32 × 128 × 128 voxels via trilinear interpolation to ensure spatial consistency.

To address class imbalance, a combined loss of binary cross-entropy and contrastive loss (weight = 0.1) was used. Data augmentation included random flipping, 90-degree rotations, ± 10% intensity perturbation, and Gaussian noise ($\sigma = 0.03$).

Statistical analysis

Differences in demographic and clinical characteristics between the training and external validation cohorts were assessed using the Mann-Whitney U test for continuous variables and the chi-squared test for categorical variables.

Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), and the DeLong test was employed to compare the AUC of M3A-Net against all baseline models. The optimal probability threshold for classification was determined by maximizing the Youden index on the internal test set, and this threshold was applied to calculate accuracy, sensitivity, and specificity across all datasets.

Model calibration was quantified using the Brier score, with a lower score indicating better calibration. Clinical utility was assessed and compared using decision curve analysis (DCA), which quantifies the net benefit across a range of probability thresholds.

All statistical analyses were performed using Python (version 3.10), with a two-sided p-value < 0.05 considered statistically significant.

Results

Study characteristics

A total of 1044 patients from four centers were enrolled in this study. The selected patients from center 1 and center 2 were merged and then randomly divided into a training cohort (n = 800, 319 patients with ECE (39.9%)) and an internal test cohort (n = 64, 20 patients with ECE (31.3%)). The patients from center 3 and center 4 were collected for external test cohorts (center 3: n = 64, 25 patients with ECE (39.1%)); center 4: n = 116, 43 patients with ECE (37.1%)). Random re-splitting showed no significant differences between the training cohort and internal test cohort in terms of age, PSA, PSAD, PI-RADS score, biopsy and surgical ISUP grade, seminal vesicle invasion (SVI), extracapsular extension (ECE), and surgical margins (p > 0.05). The clinical characteristics of the patients from the four centers are shown in Table 1.

Table 1
The baseline characteristics of training, internal and external testing cohorts

Variable	Training cohort	Internal testing cohort		External testing cohorts	
	SUH 1st (center1) and CZH 2nd (center2) (n = 864)		*P*	SUH 4th (center3, n = 64)	JNH (center 4, n = 116)
NO. of subjects	800	64		64	116
Age(years), median (IQR)	71(67–75)	69(66–76)	0.512	73(67–76)	71(67–75)
PSA level, median (IQR)	12.6(8.1–21.9)	13.5(7.3–23.4)	0.852	13.4(9.3–21.1)	13.1(8.6–27.3)
0–10 ng/mL, *n* (%)	286/800(35.7%)	25/64(39.1%)		21/64(32.8%)	39/116(33.6%)
10–20 ng/mL, *n* (%)	286/800(35.7%)	19/64(29.7%)		25/64(39.1%)	34/116(29.3%)
> 20 ng/mL, *n* (%)	228/800(28.6%)	20/64(31.2%)		18/64(28.1%)	43/116(37.1%)
PSAD ng/ml^2, median (IQR)	0.21(0.11–0.39)	0.18(0.10–0.40)	0.613	0.19(0.13–0.30)	0.22(0.13–0.42)
MRI index lesion per patient, *n* (%)			0.633		
PI-RADS 1–2	45/800(5.6%)	6/64(9.4%)		5/64(7.8%)	7/116(6.0%)
PI-RADS 3	137/800(17.1%)	7/64(10.9%)		7/64(10.9%)	17/116(14.7%)
PI-RADS 4	303/800(37.9%)	29/64(45.3%)		32/64(50.0%)	41/116(35.3%)
PI-RADS 5	315/800(39.4%)	22/64(34.4%)		20/64(31.3%)	51/116(44.0%)
Biopsy ISUP grade, *n* (%)			0.608		
ISUP 1	155/800(19.4%)	12/64(18.7%)		16/64(25.0%)	22/116(19.0%)

Notes: *PSA* Prostate-Specific Antigen; *PSAD* Prostate-Specific Antigen Density; *PCa* Prostate cancer; *ciPCa* Clinically insignificant prostate cancer; *csPCa* Clinically significant prostate cancer; *PI-RADS* Prostate Imaging Reporting and Data System; *ECE* Extracapsular extension; *SVI* Seminal vesicle invasion; *LNI* Lymph node invasion; *SM* Surgical margin; *ISUP* International Society of Urological Pathology standards.

Variable	Training cohort	Internal testing cohort		External testing cohorts	
	SUH 1st (center1) and CZH 2nd (center2) (n = 864)		*P*	SUH 4th (center3, n = 64)	JNH (center 4, n = 116)
ISUP 2	260/800(32.5%)	17/64(26.6%)		19/64(29.7%)	20/116(17.2%)
ISUP 3	202/800(25.2%)	21/64(32.8%)		14/64(21.9%)	34/116(29.3%)
ISUP 4	134/800(16.8%)	10/64(15.6%)		14/64(21.9%)	36/116(31.0%)
ISUP 5	49/800(6.1%)	4/64(6.3%)		1/64(1.5%)	4/116(3.5%)
Perineural invasion			0.752		
Present	565/800(70.6%)	44/64(68.7%)		42/64(65.6%)	80/116(69.0%)
Absent	235/800(29.4%)	20/64(31.3%)		22/64(34.4%)	36/116(31.0%)
Surgical ISUP grade, *n* (%)			0.957		
ISUP 1	39/800(4.9%)	3/64(4.7%)		5/64(7.8%)	11/116(9.4%)
ISUP 2	275/800(34.4%)	21/64(32.8%)		22/64(34.3%)	35/116(30.2%)
ISUP 3	309/800(38.6%)	26/64(40.6%)		25/64(39.1%)	45/116(38.8%)
ISUP 4	84/800(10.5%)	9/64(14.1%)		8/64(12.5%)	14/116(12.1%)
ISUP 5	93/800(11.6%)	5/64(7.8%)		4/64(6.3%)	11/116(9.5%)
Label, *n* (%)			0.903		
ciPCa	155/800(19.4%)	12/64(18.7%)		16/64(25.0%)	22/116(19.0%)
csPCa	645/800(80.6%)	52/64(81.3%)		48/64(75.0%)	94/116(81.0%)
Pathological ECE			0.174		
Present	319/800(39.9%)	20/64(31.3%)		25/64(39.1%)	43/116(37.1%)
Absent	481/800(60.1%)	44/64(68.7%)		39/64(60.9%)	73/116(62.9%)
Pathological SVI			0.866		

Notes: *PSA* Prostate-Specific Antigen; *PSAD* Prostate-Specific Antigen Density; *PCa* Prostate cancer; *ciPCa* Clinically insignificant prostate cancer; *csPCa* Clinically significant prostate cancer; *PI-RADS* Prostate Imaging Reporting and Data System; *ECE* Extracapsular extension; *SVI* Seminal vesicle invasion; *LNI* Lymph node invasion; *SM* Surgical margin; *ISUP* International Society of Urological Pathology standards.

Variable	Training cohort	Internal testing cohort		External testing cohorts	
	SUH 1st (center1) and CZH 2nd (center2) (n = 864)		*P*	SUH 4th (center3, n = 64)	JNH (center 4, n = 116)
Present	131/800(16.4%)	11/64(17.2%)		11/64(17.2%)	19/116(16.4%)
Absent	669/800(83.6%)	53/64(82.8%)		53/64(82.8%)	97/116(83.6%)
Pathological SM			0.628		
Positive	312/800(39.0%)	23/64(35.9%)		26/64(40.6%)	43/116(37.1%)
Negative	488/800(61.0%)	41/64(64.1%)		38/64(59.4%)	73/116(62.9%)
Pathological LNI			0.049		
Present	14/800(1.7%)	4/64(6.3%)		1/64(1.6%)	6/116(5.2%)
Absent	786/800(98.2%)	60/64(93.7%)		63/64(98.4%)	110/116(94.8%)

Notes: *PSA* Prostate-Specific Antigen; *PSAD* Prostate-Specific Antigen Density; *PCa* Prostate cancer; *ciPCa* Clinically insignificant prostate cancer; *csPCa* Clinically significant prostate cancer; *PI-RADS* Prostate Imaging Reporting and Data System; *ECE* Extracapsular extension; *SVI* Seminal vesicle invasion; *LNI* Lymph node invasion; *SM* Surgical margin; *ISUP* International Society of Urological Pathology standards.

Model Performance Evaluation

The quantitative performance of the proposed M3A-Net and all five baseline methods was evaluated on both the internal test set and the external validation set. The results, including AUC, accuracy, sensitivity, specificity, Brier score, and net benefit, are detailed in Table 2 and Table 3, with corresponding ROC curves, calibration curves, and performance plots visualized in Fig. 4 and Fig. 5.

Table 2
Performance Comparison on Internal Test Set

Model	AUC	Accuracy	Sensitivity	Specificity	Brier	Net Benefit
Zhao et al.	0.602	0.571	0.3704	0.7586	0.252	0.054
Hou et al.	0.679	0.571	0.04	**1**	0.361	0.018
Khosravi et al.	0.704	0.571	**1**	0.351	0.254	-0.089
Singh et al.	0.671	0.642	0.3043	0.9	0.234	0.019
Jin et al.	0.74	**0.768**	0.6	0.861	0.204	0.125
Ours	**0.873**	0.755	0.783	0.7333	**0.151**	**0.189**

Table 3
Performance Comparison on External Validation Set

Model	AUC	Accuracy	Sensitivity	Specificity	Brier	Net Benefit
hao et al.	0.655	0.628	0.377	0.815	0.233	0.056
Hou et al.	0.715	0.668	0.097	**0.966**	0.284	0.011
Khosravi et al.	0.67	0.533	**0.875**	0.306	0.249	-0.067
Singh et al.	0.811	0.722	0.833	0.667	0.195	0.056
Jin et al.	0.681	0.639	0.459	0.731	0.26	-0.022
Ours	**0.851**	**0.744**	0.867	0.683	**0.177**	**0.072**

Performance on the Internal Test Set

On the internal test set, our proposed M3A-Net demonstrated superior discriminative capability, achieving an AUC of 0.873 (Table 2), which was significantly higher than the next-best model suggesting its potential for decision support in diverse setting (ResNet-18, AUC = 0.740). The model provided a more balanced and clinically relevant trade-off with an accuracy of 0.755, sensitivity of 0.783, and specificity of 0.733, while some baseline models showed highly skewed classification behavior (Fig. 4A and 4B).

Furthermore, the model showed excellent calibration, with its prediction curve closely following the "perfect calibration" line. This was corroborated by the lowest Brier score (0.151) among all models (Fig. 4C). The combined plot (Fig. 4E) showed our model falling into the "ideal" quadrant, demonstrating strong performance in both discrimination and calibration. Decision curve analysis (Fig. 4F) suggested good potential clinical utility for our model. Across the entire range of threshold probabilities, and particularly within the clinically relevant range of 0.1 to 0.7 (Fig. 4G), M3A-Net consistently provided a higher net benefit (max NB = 0.428) compared to all five baseline models and the "treat-all" or "treat-

none" strategies. This suggests that using M3A-Net to inform surgical decisions has the potential to yield a better net clinical outcome.

Generalization on the External Validation Set

The generalization capability of our model was evaluated on the external validation set. It again achieved the highest AUC of 0.851 (Table 3), outperforming the strongest baseline (3DCNN, AUC = 0.811). The model maintained a high sensitivity of 0.867 and a specificity of 0.683 (Fig. 5A and 5B).

The model also remained well-calibrated on the external data, achieving the lowest Brier score (0.177, "good" range) (Fig. 5C). As shown in the combined plot (Fig. 5E), M3A-Net remained the top-performing model, suggesting its robustness and generalizability across different centers and devices. Decision curve analysis on the external validation set (Fig. 5F) reaffirmed the potential clinical value of our model. Even on unseen data from different centers, M3A-Net maintained the highest net benefit (max NB = 0.327) across a wide range of decision thresholds. As shown in the focused plot for the 0.1–0.7 probability range (Fig. 5G), our model consistently outperformed other methods, suggesting its potential for decision support in diverse settings.

Summary of Performance

In summary, across both internal and external multi-center datasets, the proposed M3A-Net consistently outperformed all five baseline methods. It not only demonstrated superior discriminative power (AUC 0.873 and 0.851) while maintaining a favorable sensitivity-specificity profile, but also showed better calibration with the lowest Brier scores (0.151 and 0.177). Furthermore, decision curve analysis confirmed that our model provided the highest net clinical benefit (0.189 and 0.072) in both cohorts. This strong performance in discrimination, calibration, and potential clinical utility highlights the model's promise for deployment in diverse clinical settings.

Discussion

In this study, we developed and evaluated a deep learning framework, named M3A-Net, designed for the automated preoperative prediction of ECE using memory images. Addressing challenges related to the generalizability of existing models and the variability in manual interpretation, the core objective was to develop an end-to-end solution with promising generalization capabilities. The experimental results indicate favorable performance for our model. On the internal validation set, M3A-Net achieved an AUC of 0.873, with a good balance between accuracy, sensitivity, and specificity. More importantly, on a completely independent external validation set, the model maintained a high AUC of 0.850. While sensitivity was high (0.867), the specificity was moderate (0.683). This trade-off is a critical consideration for clinical application. High sensitivity is valuable for correctly identifying patients with ECE who may require more aggressive surgery, but moderate specificity may lead to a higher number of false positives, potentially resulting in overtreatment for some patients. This discrepancy in specificity may be attributed to the difficulty in distinguishing ECE from periprostatic fibrosis or inflammation on

MRI—a challenge even for experienced radiologists. Future work should focus on optimizing this balance through the integration of more granular clinical data.

A significant strength of M3A-Net, beyond its discriminative power, is its excellent calibration and potential clinical utility. While high AUC values confirm the model's ability to distinguish ECE-positive from ECE-negative cases, our analysis demonstrated that its reliability is also strong. The model achieved the lowest Brier scores on both internal (0.151, "excellent") and external (0.177, "good") datasets, indicating that its predicted probabilities are reliable. This level of calibration is critical for clinical adoption. Furthermore, the decision curve analysis suggested that our model provided the highest net benefit across a wide range of practical thresholds in both cohorts (0.189 and 0.072). This result suggests its potential utility as a decision-support tool. It is important to note, however, that DCA is a theoretical analysis of clinical benefit. The true impact of the model on clinical practice can only be ascertained through prospective validation studies that assess its influence on actual surgical decisions and patient outcomes.

The success of our model can be attributed to several key design choices: first, the use of independent, parallel 3D residual networks to process T2WI, ADC, and DWI modalities separately. Second, the introduction of channel attention and contrastive learning mechanisms. Finally, an efficient multi-modal fusion layer organically combines information from different sources. Furthermore, this study utilized an automated 3D prostate segmentation technique (nnU-Net2), integrating the entire diagnostic workflow. This not only reduces the workload for radiologists but also may decrease the subjectivity introduced by manual delineation.

Despite the encouraging results, this study has several important limitations.

First, the study is retrospective in nature. Although it includes multi-center data, this design has inherent biases. Future research should include prospective, multi-center, multi-national clinical trials to further validate its effectiveness and reliability in real-world clinical settings before any consideration for clinical deployment.

Second, a significant limitation is the lack of a direct comparison against radiologist performance. To establish the clinical value of an AI model, it is essential to demonstrate that it can perform comparably to, or improve upon, the current standard of care. This comparison is a critical next step for future validation.

Third, while the workflow incorporates an automated segmentation model (nnU-Net2) to create an end-to-end pipeline, this segmentation component was not independently validated on the external datasets. The performance of the upstream segmentation can impact the downstream prediction task, and thus, its generalizability requires separate verification.

Fourth, the model currently relies only on imaging data and does not yet integrate important clinical indicators such as Prostate-Specific Antigen (PSA) levels or Gleason scores. Combining these non-

imaging data with deep learning features could lead to a more comprehensive, multi-dimensional prediction model, further enhancing predictive accuracy.

Lastly, the "inherent opacity of deep learning models" nature of the model makes its decision-making process difficult to interpret. Future work could explore eXplainable AI (XAI) techniques, such as generating saliency maps, to visualize the key image regions the model focuses on when making a diagnosis, thereby increasing clinicians' trust in the model's outputs.

Conclusion

This study developed and evaluated an automated deep learning framework, M3A-Net, for the preoperative assessment of ECE status. The model demonstrated promising performance in discrimination, calibration, and potential clinical utility, while maintaining good generalization capabilities across a multi-center, multi-device dataset. As a potential decision-support tool, M3A-Net shows promise to assist clinicians in optimizing surgical planning. However, this framework represents a promising step towards more precise treatment but requires substantial further validation—including prospective studies and comparisons with radiologists—before its clinical value can be determined and it can be considered for clinical implementation.

Abbreviations

ADC	Apparent diffusion coefficient
ADT	Androgen deprivation therapy
AI	Artificial intelligence
AUC	Area under the curve
BPH	Benign prostatic hyperplasia
CsPCa	Clinically significant prostate cancer
CiPCa	Clinically insignificant prostate cancer
DCA	Decision curve analysis
DCE	Dynamic contrast enhanced
DWI	Diffusion weighted imaging
ECE	Extracapsular extension
FDA	Food and Drug Administration
ISUP	International Society of Urological Pathology
mpMRI	Multiparametric magnetic resonance imaging
NCI	National Cancer Institute
NPV	Negative predictive value
Pca	Prostate cancer
PI-RADS	Prostate Imaging Reporting and Data System
PPV	Positive predictive value
PSA	Prostate specific antigen
PSM	Positive surgical margin
ROC	Receiver operating characteristic
RP	Radical prostatectomy
ROI	Region of interest
T2WI	T2 weighted imaging

Declarations

Ethics approval and consent to participate

This multicenter study was approved by the ethics committee of the First Affiliated Hospital of Soochow University (No. 2025021). Due to the retrospective nature of the study, the requirement for informed

consent was waived. This study was registered with the Chinese Clinical Trial Registry (ChiCTR) (registration number: ChiCTR2500105361; registration date: July 2, 2025). All procedures conducted in the studies involving human participants were in accord with the 1964 Helsinki Declaration and its later amendments.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analyzed during the current study are not publicly available due to data privacy and intellectual property issues but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was funded by The National Natural Science Foundation of China (No. 82002715), The Suzhou science and technology development plan project (SKY2023154) and Jiangsu Health International Exchange Program.

Authors' contributions

Z.X.L. and Y.Y.M. conceived, designed, supervised and sponsored the project; B.L.Z., Z.K.Z., D.Y.W., H.Y.Z., and L.F.Z. collected and pre-processed all data and performed the research; S.F.L. and Z.X.L. performed imaging annotation, clinical data review and proposed the model, B.L.Z. drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

During the preparation of this work, the authors used AI-assisted tools in order to improve the language quality and readability. After using this tool, the authors reviewed and edited the content and take full responsibility for the content of the publication.

Authors' information

Baolin Zhang, Zhenkun Zhang, Zhixin Ling: Department of Urology, The First Affiliated Hospital of Soochow University; No.899 Pinghai Road, Suzhou, 215031，Jiangsu Province, China;

Danyu Wu, Department of Urology, The Third People's Hospital of Xiangcheng District, No.568, Weizhong Road, Suzhou, 215134, China;

Haoyang Zhang: Department of Urology, The Fourth Affiliated Hospital of Soochow University; No.9, Chongwen Road, 215000, China;

Shanfu Lu: Jiaxing Yaohe Technology Co., Ltd; No. 138, Xiuzhou Avenue, Jiaxing, 314031，Zhejiang Province, China;

Lifeng Zhang: Department of Urology, Changzhou Second People's Hospital, Changzhou Medical Center, Changzhou 213000, China;

Yuanyuan Mi: Department of Urology, Affiliated Hospital of Jiangnan University, 1000 Hefeng Road, Wuxi 214122, China;

References

1. 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):340-51. doi: 10.1016/j.eururo.2019.02.033.
2. Yossepowitch O, Briganti A, Eastham JA, et al. Positive surgical margins after radical prostatectomy: a systematic review and contemporary update. Eur Urol. 2014;65(2):303-13. doi: 10.1016/j.eururo.2013.07.039.
3. Nguyen PL, Je Y, Schutz FA, et al. Association of androgen deprivation therapy with cardiovascular death in patients with prostate cancer: a meta-analysis of randomized trials. JAMA. 2011;306(21):2359-66. doi: 10.1001/jama.2011.1745.
4. Nguyen PL, Alibhai SM, Basaria S, et al. Adverse effects of androgen deprivation therapy and strategies to mitigate them. Eur Urol. 2015;67(5):825-36. doi: 10.1016/j.eururo.2014.07.010.
5. Dearnaley D, Syndikus I, Sumo G, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: preliminary safety results from the CHHiP randomised controlled trial. Lancet Oncol. 2012;13(1):43-54. doi: 10.1016/S1470-2045(11)70293-5.
6. Pinkawa M, Asadpour B, Piroth MD, et al. Health-related quality of life after permanent I-125 brachytherapy and conformal external beam radiotherapy for prostate cancer--a matched-pair comparison. Radiother Oncol. 2009;91(2):225-31. doi: 10.1016/j.radonc.2008.11.005.
7. Suardi N, Moschini M, Gallina A, et al. Nerve-sparing approach during radical prostatectomy is strongly associated with the rate of postoperative urinary continence recovery. BJU Int. 2013;111(5):717-22. doi: 10.1111/j.1464-410X.2012.11315.x.
8. Maruo M, Goto Y, Miyazaki K, et al. Novel nerve-sparing robot-assisted radical prostatectomy with endopelvic fascia preservation and long-term outcomes for a single surgeon. Sci Rep. 2024;14(1):926. doi: 10.1038/s41598-024-51598-3.

9. Pellegrino F, Falagario UG, Knipper S, et al. Assessing the Impact of Positive Surgical Margins on Mortality in Patients Who Underwent Robotic Radical Prostatectomy: 20 Years' Report from the EAU Robotic Urology Section Scientific Working Group. Eur Urol Oncol. 2024;7(4):888-96. doi: 10.1016/j.euo.2023.11.021.
10. Mehralivand S, Shih JH, Harmon S, Smith C, Bloom J, Czarniecki M, et al. A Grading System for the Assessment of Risk of Extraprostatic Extension of Prostate Cancer at Multiparametric MRI. Radiology. 2019;290(3):709-19. doi: 10.1148/radiol.2018181278.
11. Falagario UG, Jambor I, Ratnani P, et al. Performance of prostate multiparametric MRI for prediction of prostate cancer extra-prostatic extension according to NCCN risk categories: implication for surgical planning. Minerva Urol Nefrol. 2020;72(6):746-54. doi: 10.23736/S0393-2249.20.03688-7.
12. Ruprecht O, Weisser P, Bodelle B, Ackermann H, Vogl TJ. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. Eur J Radiol. 2012;81(3):456-60. doi: 10.1016/j.ejrad.2010.12.076.
13. Lin Y, Belue MJ, Yilmaz EC, et al. Deep learning-based image quality assessment: impact on detection accuracy of prostate cancer extraprostatic extension on MRI. Abdom Radiol (NY). 2024;49(8):2891-901. doi: 10.1007/s00261-024-04468-5.
14. Khosravi P, Saikali S, Alipour A, et al. AutoRadAI: a versatile artificial intelligence framework validated for detecting extracapsular extension in prostate cancer. Biol Methods Protoc. 2025;10(1):bpaf032. doi: 10.1093/biomethods/bpaf032.
15. Priester A, Mota SM, Grunden KP, et al. Extracapsular extension risk assessment using an artificial intelligence prostate cancer mapping algorithm. BJUI Compass. 2024;5(10):986-97. doi: 10.1002/bco2.421.
16. Moroianu SL, Bhattacharya I, Seetharaman A, et al. Computational Detection of Extraprostatic Extension of Prostate Cancer on Multiparametric MRI Using Deep Learning. Cancers (Basel). 2022;14(12). doi: 10.3390/cancers14122821.
17. Guerra A, Orton MR, Wang H, et al. Clinical application of machine learning models in patients with prostate cancer before prostatectomy. Cancer Imaging. 2024;24(1):24. doi: 10.1186/s40644-024-00666-y.
18. Guerra A, Alves FC, Maes K, et al. Early biomarkers of extracapsular extension of prostate cancer using MRI-derived semantic features. Cancer Imaging. 2022;22(1):74. doi: 10.1186/s40644-022-00509-8.
19. Hou Y, Zhang YH, Bao J, et al. Artificial intelligence is a promising prospect for the detection of prostate cancer extracapsular extension with mpMRI: a two-center comparative study. Eur J Nucl Med Mol Imaging. 2021;48(12):3805-16. doi: 10.1007/s00259-021-05381-5.
20. van den Berg I, Soeterik TFW, van der Hoeven E, et al. The Development and External Validation of Artificial Intelligence-Driven MRI-Based Models to Improve Prediction of Lesion-Specific Extraprostatic Extension in Patients with Prostate Cancer. Cancers (Basel). 2023;15(22). doi: 10.3390/cancers15225452.

21. Epstein JI, Egevad L, Amin MB, 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.
22. Zhao L, Bao J, Wang X, et al. Detecting adverse pathology of prostate cancer with a deep learning approach based on a 3D swin-transformer model and biparametric MRI: a multicenter retrospective study. J Magn Reson Imaging. 2024;59(6):2101-12. doi: 10.1002/jmri.28963.
23. Singh SK, Sinha A, Singh H, et al. A novel deep learning-based technique for detecting prostate cancer in MRI images. Multimedia Tools and Applications. 2024;83:14173–87. doi: 10.1007/s11042-023-15793-0.
24. Jin L, Yu Z, Gao F, Li M. T2-weighted imaging-based deep-learning method for noninvasive prostate cancer detection and Gleason grade prediction: a multicenter study. Insights Imaging. 2024;15(1):111. doi: 10.1186/s13244-024-01682-z.

Figures

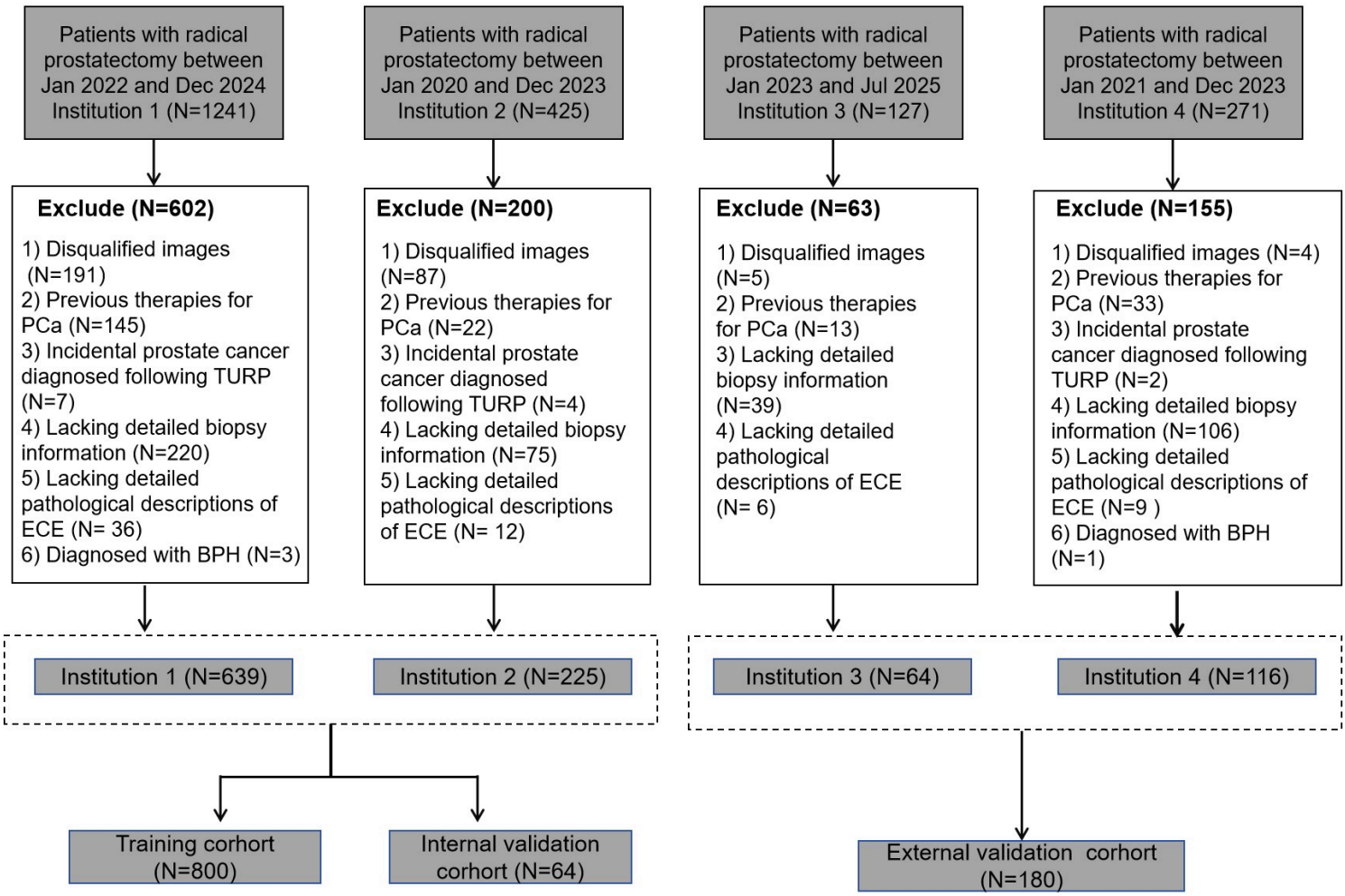


Figure 1

The study flow chart. Notes: Center 1, SUH 1st, the First Affiliated Hospital of Soochow University，center 2, CZH 2nd, the Second People's Hospital of Changzhou; center 3, SUH 4th, The Fourth Affiliated Hospital of Soochow University; center 4, JUH, the Affiliated Hospital of Jiangnan University; PCa, prostate cancer; TURP: transurethral resection of the prostate; ECE: extracapsular extension.

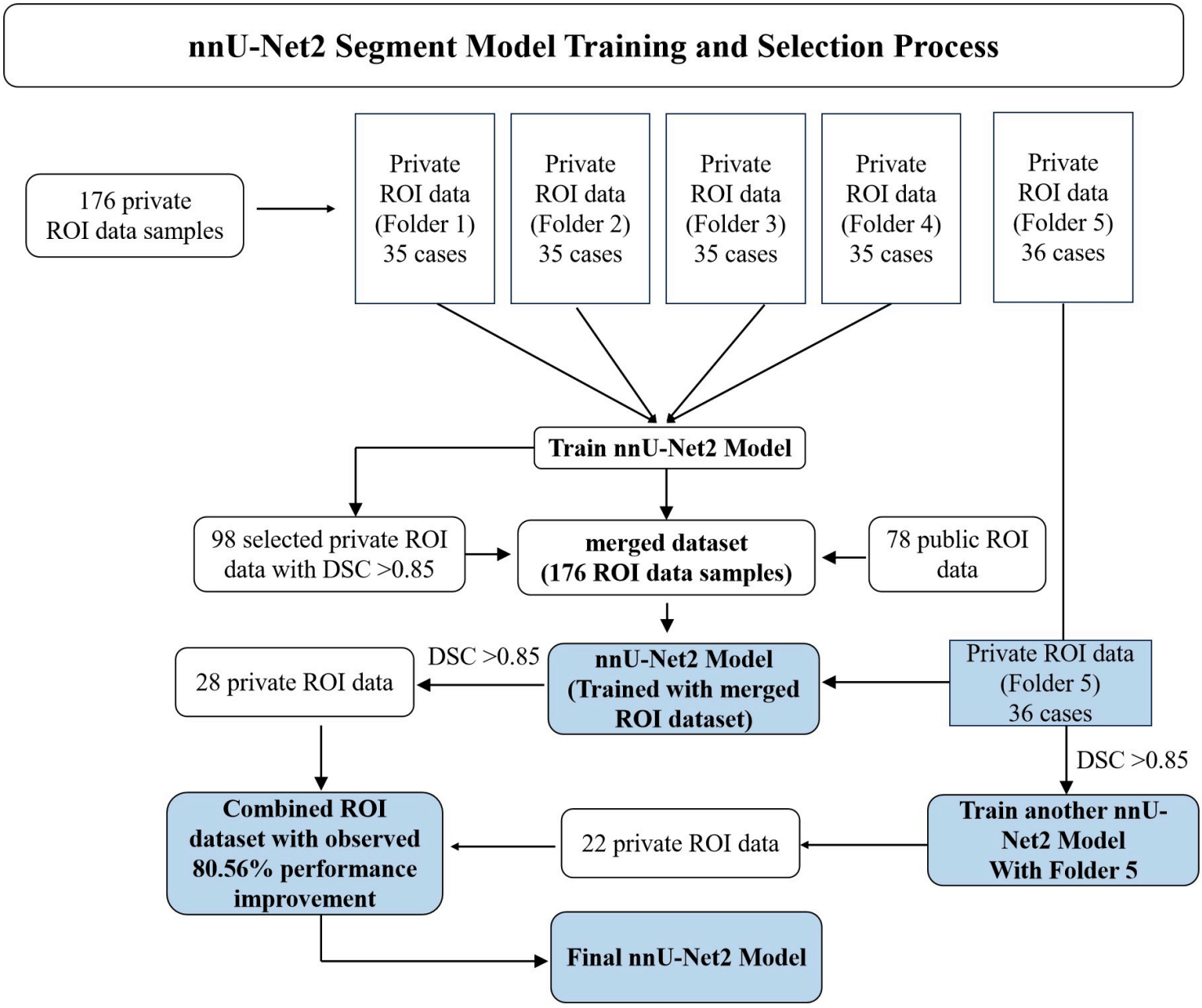


Figure 2

nnU-Net2 model training and selection pipeline.

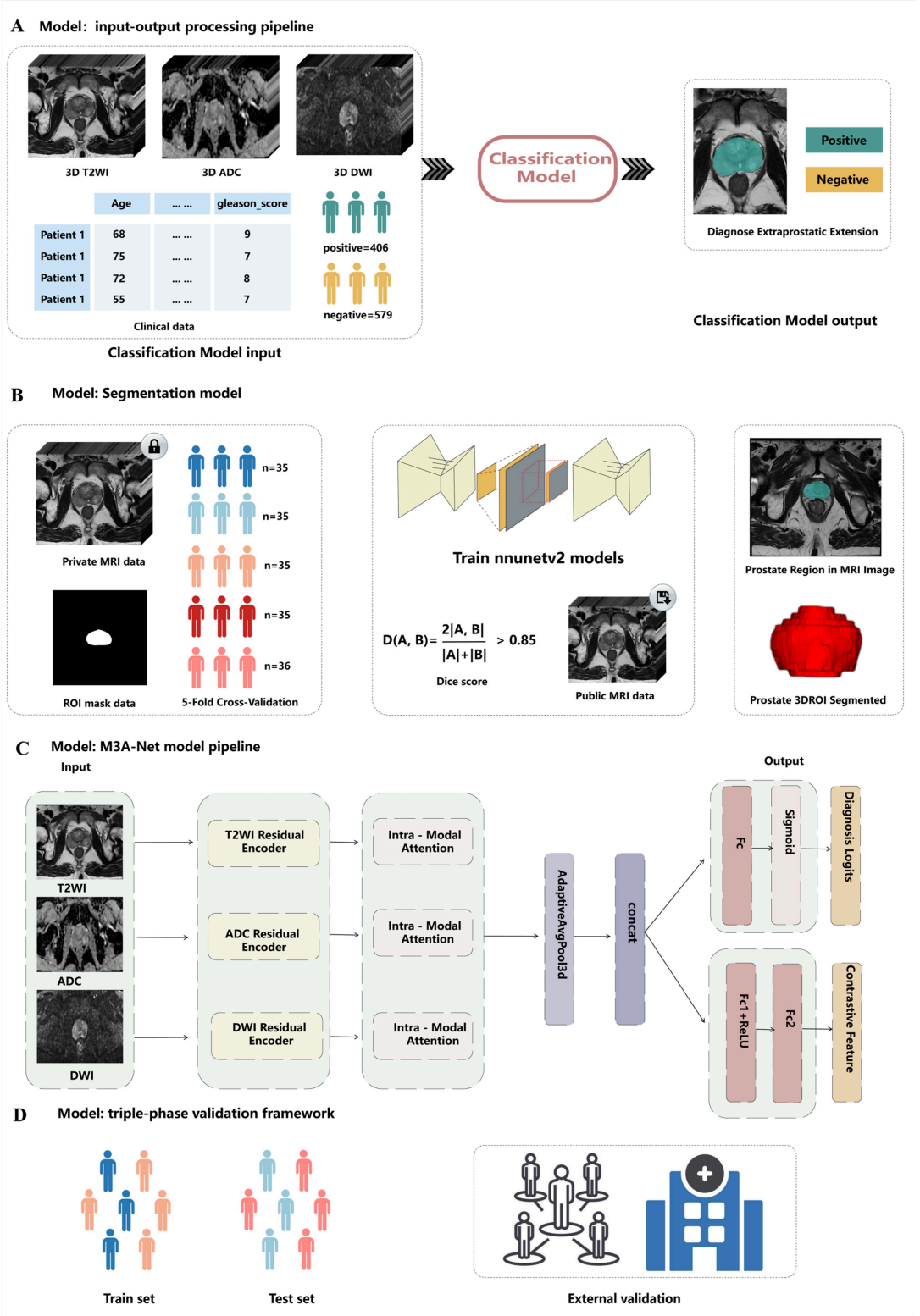


Figure 3

Overview of the Proposed Framework. The framework is composed of four main stages: (A) the input-output processing pipeline, (B) the automated 3D prostate segmentation model, (C) an overview of the proposed M3A-Net architecture, and (D) the triple-phase validation framework. The following sections provide a detailed description of each component.

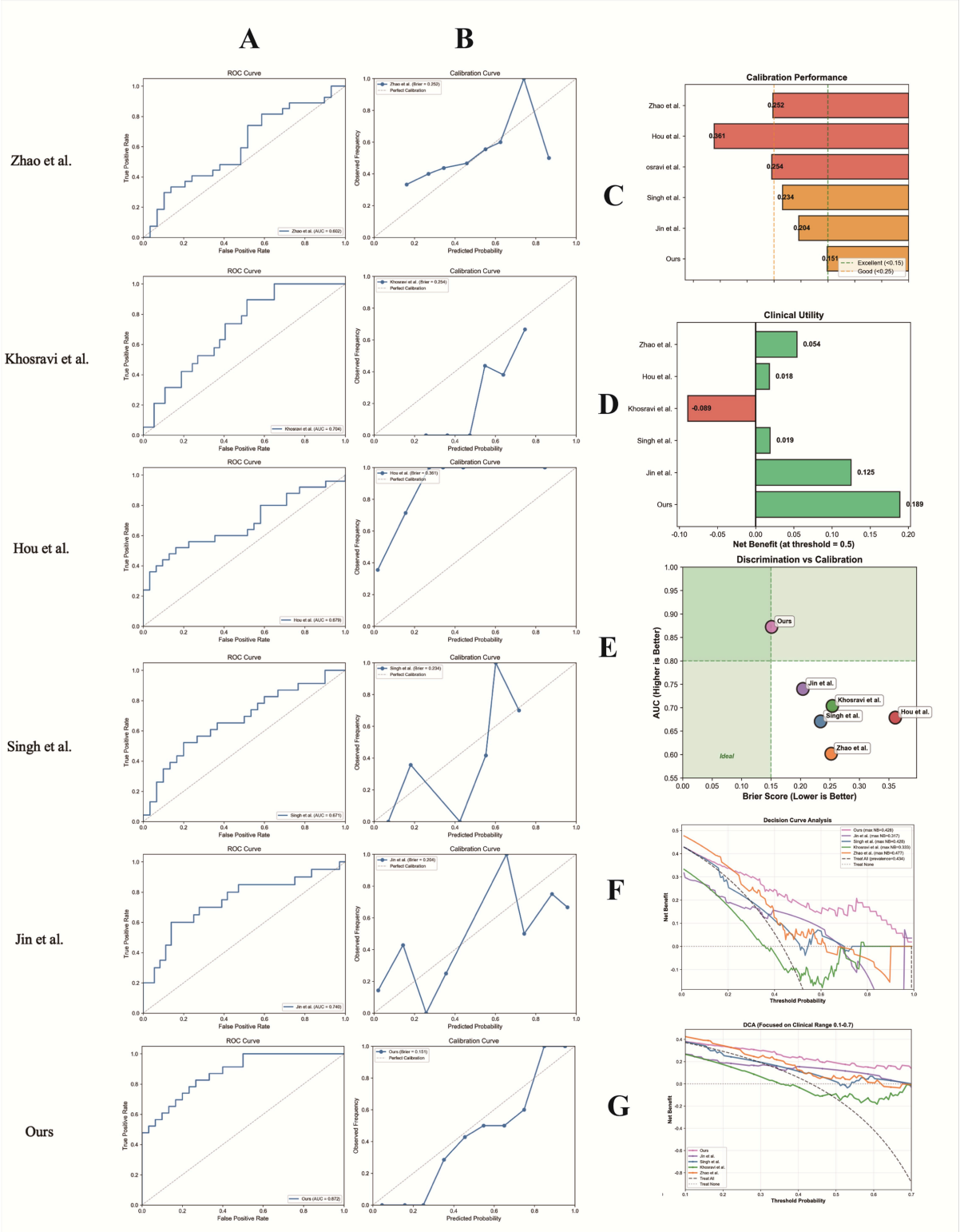


Figure 4

Performance comparison of six deep learning models on the internal test set. (A-B) Receiver Operating Characteristic (ROC) curves and calibration curves for ECE prediction, comparing our proposed method against five state-of-the-art (SOTA) methods on the internal test set. (C) Calibration performance (Brier score, lower is better); (D) Clinical utility (net benefit, higher is better); (E) Relationship between discrimination and calibration. The Ours model demonstrates superior performance across all metrics

(AUC=0.873, Brier=0.151, Net Benefit=0.189). (F-G) Detailed decision curve analysis (DCA) on the internal test set. The upper panel shows the full probability range, while the lower panel focuses on the clinically relevant range of 0.1 to 0.7. M3A-Net achieved the highest maximum net benefit (0.428).

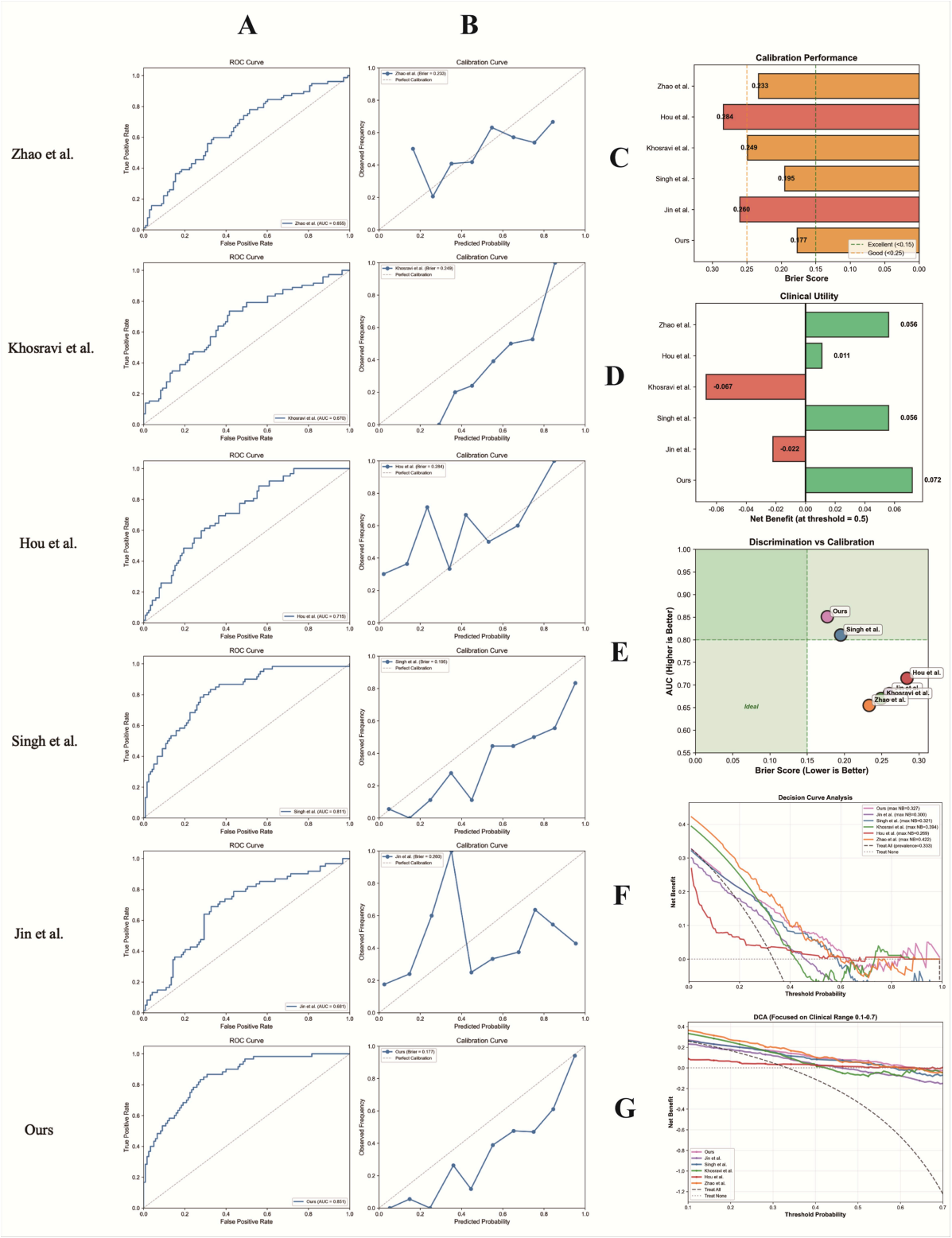


Figure 5

Performance comparison on the external validation set. (A-B) Receiver Operating Characteristic (ROC) curves and calibration curves for ECE prediction, comparing our proposed method against five state-of-

the-art (SOTA) methods on the external validation set. (C) Calibration performance (Brier score); (D) Clinical utility overview; (E) Discrimination vs. Calibration plot. M3A-Net maintains the best comprehensive performance (AUC=0.851, Brier=0.177). (F-G) Detailed decision curve analysis (DCA) on the external validation set. The upper panel displays the net benefit across all thresholds, and the lower panel focuses on the 0.1–0.7 range. M3A-Net maintained superior clinical utility (max net benefit = 0.327), validating its generalizability.

Supplementary Files

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

- Additionalfile1.docx