Predicting Prostate Cancer Upgrading of Biopsy Gleason Grade Group at Radical Prostatectomy Using Machine Learning-Assisted Decision-support Models

Hailang Liu
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Kun Tang
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Ejun Peng
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Liang Wang
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Ding Xia
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Zhiqiang Chen (lhl20180129@163.com)
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
https://orcid.org/0000-0002-0828-7689

Research

Keywords: Prostate cancer, biopsy cores, Gleason grade group, upgrading, machine learning

Posted Date: July 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42727/v1

Version of Record: A version of this preprint was published at Cancer Management and Research on December 1st, 2020. See the published version at https://doi.org/10.2147/CMAR.S286167.

Abstract

Objective: To develop a machine learning (ML)-assisted model capable of accurately predicting the probability of biopsy Gleason grade group upgrading before making treatment decisions by integrating multiple clinical characteristics.

Materials and Methods: We retrospectively collected data from PCa (prostate cancer) patients who underwent systematic biopsy and radical prostatectomy from January 2015 to December 2019 at Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology. The study cohort was divided into training and testing datasets in a 70:30 ratio for further analysis. Four ML-assisted models were developed from 16 clinical features using logistic regression (LR), logistic regression optimized by Lasso regularization (Lasso-LR), random forest (RF) and support vector machine (SVM). The area under the curve (AUC) was applied to determine the model with the highest discrimination. Calibration plots were used to investigate the extent of over- or underestimation of predicted probabilities relative to the observed probabilities in models.

Results: In total, 530 PCa patients were included, with 371 patients in the training dataset and 159 patients in the testing dataset. The Lasso-LR model showed good discrimination with an AUC, accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of 0.776, 0.712, 0.679, 0.745, 0.730 and 0.695, respectively, followed by SVM (AUC 0.740, 95%CI: 0.690–0.790), LR (AUC 0.725, 95%CI: 0.674–0.776) and RF (AUC 0.666, 95%CI: 0.618–0.714). Validation of the model showed that the Lasso-LR model had the best discriminative power (AUC 0.735, 95%CI: 0.656–0.813), followed by SVM (AUC 0.723, 95%CI: 0.644–0.802), LR (AUC 0.697, 95%CI: 0.615–0.778) and RF (AUC 0.607, 95%CI: 0.531–0.684) in the testing dataset. Both the Lasso-LR and SVM models were well-calibrated.

Conclusions: The Lasso-LR model had good discrimination in the prediction of patients at high risk of harboring incorrect Gleason grade group assignment, and the use of this model may be greatly beneficial to urologists in treatment planning, patient selection, and the decision-making process for PCa patients.

Introduction

Despite first being introduced in 1966, the Gleason score (GS) remains the most widely used grading system for prostate cancer (PCa) [1]. The GS grading system was updated in 2005 and 2014 and a new five-tier grade group (GG) system was proposed and developed: GG 1 (GS $\leq$ 6), GG 2 (GS 3 + 4 = 7), GG 3 (GS 4 + 3 = 7), GG 4 (GS 8), and GG 5 (GS 9 and 10) [2, 13]. Appropriate clinical management for PCa patients depends on accurate risk stratification, which is mainly reliant on the pretreatment prostate-specific antigen (PSA) level, Gleason grade group (GG) of positive biopsy cores and tumor stage. However, up to 56% of high-grade patients at initial biopsy tend to be overestimated, compared with prostatectomy specimens due to the sampling error of biopsy and the multifocal nature of PCa [3, 4]. In addition, some low risk patients who embark on active surveillance (AS) will be upgraded to higher grade at RP so they are not suitable candidates for AS [5, 6]. The discordance between initial biopsy GG and radical [illegible] ct over- and under-treatment [7]. Therefore, for the

management of PCa, it is of pivotal importance to identify PCa patients at a higher risk of upgrading at RP before making treatment-related decisions.

Machine learning (ML) is a branch of artificial intelligence that trains computers to perform tasks by observing patterns in large datasets and using them to derive rules or mathematical algorithms that optimize task performance regarding making predictions or decisions [8]. Owing to the ability of ML algorithms to improve the accuracy of predicting diseases and subsequent outcomes over the use of traditional statistical models, they have been applied extensively in the field of clinical research [9, 10]. In the present study, we apply machine learning algorithms to the dataset, with the goal of identifying those patients at high risk of harboring upgrading at RP before making treatment decisions, and to determine the best predictive models.

Additionally, there is still no consensus on how to choose a ‘case level’ biopsy GG (GS) for patients regarding the reporting of the ‘worst/highest’ and ‘global/overall’ GG (GS) [7, 11]. Considering that the biopsy global GG is more likely be in line with RP GG, we selected global GG together with other preoperative clinical parameters to construct predictive models to calculate the probability of upgrading for each PCa patient.

Materials And Methods

Patient selection and study parameters

Patients who underwent radical prostatectomy at Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology between January 2015 and December 2019 were retrospectively enrolled in this study. Data collection was approved by the institutional review board. The inclusion criteria were as follows: (1) multi-parametric MRI (mp-MRI) performed for all patients before surgery; (2) standard systematic (12-core) transrectal ultrasonography (TRUS)-guided biopsy prior to surgery performed for all patients; and (3) final pathological results of each patient including a detailed description of Gleason grade group. The following exclusion criteria were applied: (1) neoadjuvant therapy prior to MRI examination; (2) patients with incomplete clinical data; and (3) MRI images of unsatisfactory quality. The clinical parameters included patient age, total prostate-specific antigen (TPSA), prostate volume (PV = Height × Width × Length × 0.52), PSA density (PSAD), %fPSA (free PSA/TPSA), maximum diameter of index lesion (D-max), the Prostate Imaging Reporting and Data System (PI-RADS) score, clinical and pathological T stage (T1-2, T3a, T3b and T4), apical involvement at MRI, global biopsy GG, number of positive cores, number of cores with clinically significant PCa (csPCa, defined as cores with GG $\geq$ 2), maximum tumor length at biopsy core, percentage of tumor in total biopsy cores and GG at RP. The MRI findings were re-reported and scored by the same dedicated radiologist on a five-point scale using the modified PI-RADS version 2 criteria [12]. If multiple tumor foci existed on the MRI, only the highest PI-RADS score of index lesions and the maximum diameter of the largest tumor were included in the analysis. GG of the biopsy specimen was assigned following the 2014 ISUP criteria [13]. The global GG of the biopsy

was defined as the most prevalent GG among all positive cores. The upgrades from biopsy to RP represented at least one grade difference in the GG.

Development, Validation, And Performance Of Ml-based Models

The dataset was randomly split into 2 datasets: 70% for model training and 30% for model testing. For model training, data from the training set were used to approximate model parameters. Four ML algorithms were performed to build predictive models: logistic regression (LR), logistic regression optimized by Lasso regularization (Lasso-LR), random forest classifier (RF) and support vector machine (SVM) integrated with recursive feature elimination (RFE).

Logistic regression is one of the most common ML algorithms for the classification of binary outcomes. We performed univariate and multivariable logistic regression analysis to investigate the association between clinical variables and upgrading at RP. In addition, according to the results of multivariable analysis, we selected significant predictors ($P < 0.05$) and their corresponding coefficients to construct the predictive model. The LR model was derived from the following formula:

$$Y = \text{Intercept} + \sum_{i=1}^{n} \beta\text{i} \times \log(xi), (1)$$

where Y is the output, βi is the nonzero coefficient, and xi is the selected clinical features based on results of multivariable logistic regression analysis [14].

The least absolute shrinkage and selection operator (Lasso) is a popular ML algorithm with outstanding feature selection capability. The LASSO preferentially shrinks some predictor coefficients to zero by penalizing the absolute values of the regression coefficients [15, 16]. In this study, the optimized logistic regression coefficients were estimated given a boundary (“L1 Norm”) to the sum of absolute standardized regression coefficients [15, 16]. The Lasso-LR model was also derived from the formula (1).

Random forest is an ensemble learning method that performs classification or regression by combining the voting results of multiple decision trees, which has been extensively employed in the fields of clinical research and bioinformatics [17, 18]. Bootstrap aggregation, which is also called bagging, is the core of RF algorithms. Each decision tree is trained on randomly sampled subsets in the training data, while sampling is undertaken with the replacement. The final RF model is constructed based on the majority vote results from individually developed decision trees in the forest.

Support vector machine is a supervised learning model with an associated learning algorithm that analyzes data used for classification and regression [19]. The objective of applying SVM is to find the best line in two dimensions or the best hyperplane in more than two dimensions to help the space be separated into classes [19]. In the present study, recursive feature elimination was integrated with the SVM classifier

training, and the SVM model training was based on the use of a linear kernel. RFE was initially proposed to enable SVM to perform feature selection by iteratively training a model, ranking features, and then removing the lowest ranking features [20]. The iteration was repeated until the desired number of features was reached. By adding the ranked features returned by the SVM one by one from most to least important, we eventually selected parameters that produced the greatest accuracy and the lowest average error.

In order to acquire the probability of upgrading in the four predictive models, we then converted output values of models to the probabilities (P_i) by employing a sigmoid function:

$$P_i = 1/(1 + \exp(-Y)), (2)$$

where Y is the output value of predictive models, and P_i indicates the probabilities of harboring upgrading at RP [14].

Model evaluation was carried out by examining discrimination and calibration. The receiver operating characteristic (ROC) curve analysis was used to evaluate the discrimination ability of predictive models in both the training dataset and testing dataset; the discrimination ability of each model was quantified by the area under the ROC curve (AUC). Moreover, discrimination metrics including accuracy, sensitivity, specificity, Youden index (YI), positive predictive value (PPV) and negative predictive value (NPV) were also applied to assess the discriminative power of predictive models. Comparisons between ROC Curves were performed using the method described by DeLong et al. [21]. As logistic regression analysis was one of the most widely used statistical methods, we used the LR model as the reference in the pairwise comparison of AUC. Calibration plots were used to investigate the extent of over- or under-estimation of predicted probabilities relative to the observed probabilities.

Statistical analyses were performed using R software (Version 3.6.0; https://www.R-project.org) with the following packages: ‘rms’, ‘glmnet’, ‘caret’, ‘rpart’, ‘randomForest’, ‘gplots’, ‘e1071’, ‘kernlab’, ‘pROC’ and ‘MachineShop’. $P < 0.05$ was considered statistically significant.

Results

Baseline characteristics and pathological results

Table 1 lists patient characteristics and pathological results in the total population (n = 530). The median patient age of the overall cohort was 69 (IQR: 63–75) years. The median TPSA value was 21.0 (IQR: 10.9–42.2) ng/ml. The median D-max on MRI was 1.9 (IQR: 1.3–2.7) cm. Most patients had clinical stage T1-2 (57.9%, n = 307) and PI-RADS score 5 (58.6%, n = 310). Table 4 details the concordance between the biopsy global GG and the final RP GG, and the corresponding downgrades and upgrades for GG 1–4. The most 1 (31.3%, n = 166) and GG 2 (26.0%, n = 138). Overall, 262

patients (49.4%) experienced upgrading at final pathology. The overall incidence of biopsy GG 1 upgrading was 120 (72.3%) of 166 patients, of which most were to GG 2 (50.0%, n = 83), followed by GG 3 (12.7%, n = 21), GG 4 (6.6%, n = 11) and GG 5 (3.0%, n = 5). Biopsy GG 3 (47.7%) and GG 4 (44.3%) showed the highest agreements when compared with RP GG. Patients with lower biopsy GG were more likely to harbor upgrading at RP. Table 2 and Table 3 summarize patient characteristics and pathological results of the training dataset and testing dataset, respectively.

Table 1
Characteristics of total population

	Overall (n = 530)
Age (y), median (IQR)	69 (63–75)
PV (ml), median (IQR)	37.5 (29.7–49.3)
D-max (cm), median (IQR)	1.9 (1.3–2.7)
TPSA (ng/ml), median (IQR)	21.0 (10.9–42.2)
%fPSA (fPSA/TPSA) (n, %)	
≤ 0.16	423 (79.8)
> 0.16	107 (20.2)
PSAD (n, %)	
≤ 0.20	70 (13.2)
> 0.20	460 (86.8)
Clinical T stage at MRI (n, %)	
T1-2	307 (57.9)
T3a	75 (14.2)
T3b	140 (26.4)
T4	8 (1.5)
PI-RADS score (n, %)	
1–2	24 (4.5)
3	67 (12.6)
4	129 (24.3)
5	310 (58.6)
Apical involvement at MRI (n, %)	
Yes	301 (56.8)
No	229 (43.2)
Biopsy grade group (n, %)	
1	166 (31.3)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: Prostate Imaging Reporting and Data System.

	Overall (n = 530)
2	138 (26.0)
3	111 (20.9)
4	115 (21.8)
No. of positive biopsies, median (IQR)	5 (3–9)
Presence of cores with csPCa, (n, %)	
Yes	433 (81.7)
No	97 (18.3)
Presence of cores with tumor length ≥ 0.6 cm, median (IQR)	
Yes	402 (75.8)
No	128 (24.2)
Maximum tumor length in single core (cm), median (IQR)	0.9 (0.6–1.4)
Total tumor length of positive cores (cm), median (IQR)	2.7 (1.2–5.4)
Percentage of tumor in total biopsy cores (%), median (IQR)	18.7 (7.9–36.0)
Presence of upgrading at final pathology (n, %)	
Yes	262 (49.4)
No	268 (50.6)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: the Prostate Imaging Reporting and Data System.

Table 2
Characteristics of the training dataset

	Overall (n = 371)
Age (y), median (IQR)	69 (63–75)
PV (ml), median (IQR)	37.1 (30.0-47.8)
D-max (cm), median (IQR)	1.9 (1.3–2.7)
TPSA (ng/ml), median (IQR)	20.9 (10.9–42.4)
%fPSA (fPSA/TPSA) (n, %)	
≤ 0.16	293 (79.0)
> 0.16	78 (21.0)
PSAD (n, %)	
≤ 0.20	49 (13.2)
> 0.20	322 (86.8)
Clinical T stage at MRI (n, %)	
T1-2	218 (58.8)
T3a	50 (13.5)
T3b	100 (27.0)
T4	3 (0.7)
PI-RADS score (n, %)	
1–2	20 (5.4)
3	47 (12.7)
4	90 (24.3)
5	214 (57.6)
Apical involvement at MRI (n, %)	
Yes	207 (55.8)
No	164 (44.2)
Biopsy grade group (n, %)	
1	118 (31.8)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: Prostate Imaging Reporting and Data System.

	Overall (n = 371)
2	98 (26.4)
3	76 (20.5)
4	79 (21.3)
No. of positive biopsies, median (IQR)	5 (3–9)
Presence of cores with csPCa, (n, %)	
Yes	301 (81.1)
No	70 (18.9)
Presence of cores with tumor length ≥ 0.6 cm, median (IQR)	
Yes	290 (78.2)
No	81 (21.8)
Maximum tumor length in single core (cm), median (IQR)	0.9 (0.6–1.4)
Total tumor length of positive cores (cm), median (IQR)	2.8 (1.3–5.6)
Percentage of tumor in total biopsy cores (%), median (IQR)	17.3 (8.3–35.9)
Presence of upgrading at final pathology (n, %)	
Yes	187 (50.4)
No	184 (49.6)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: the Prostate Imaging Reporting and Data System.

Table 3
Characteristics of the testing dataset

	Overall (n = 159)
Age (y), median (IQR)	69 (62–73)
PV (ml), median (IQR)	38.2 (28.5–52.0)
D-max (cm), median (IQR)	1.8 (1.3–2.7)
TPSA (ng/ml), median (IQR)	21.4 (10.7–41.0)
%fPSA (fPSA/TPSA) (n, %)	
≤ 0.16	130 (81.8)
> 0.16	29 (18.2)
PSAD (n, %)	
≤ 0.20	21 (13.2)
> 0.20	138 (86.8)
Clinical T stage at MRI (n, %)	
T1-2	89 (56.0)
T3a	25 (15.7)
T3b	40 (25.2)
T4	5 (3.1)
PI-RADS score (n, %)	
1–2	4 (2.5)
3	20 (12.6)
4	39 (24.5)
5	96 (60.4)
Apical involvement at MRI (n, %)	
Yes	94 (59.1)
No	65 (40.9)
Biopsy grade group (n, %)	
1	48 (30.2)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: Prostate Imaging Reporting and Data System.

	Overall (n = 159)
2	40 (25.2)
3	35 (22.0)
4	36 (22.6)
No. of positive biopsies, median (IQR)	6 (3–9)
Presence of cores with csPCa, (n, %)	
Yes	132 (83.0)
No	27 (17.0)
Presence of cores with tumor length ≥ 0.6 cm, median (IQR)	
Yes	112 (70.4)
No	47 (29.6)
Maximum tumor length in single core (cm), median (IQR)	0.9 (0.5–1.3)
Total tumor length of positive cores (cm), median (IQR)	2.7 (1.0-5.2)
Percentage of tumor in total biopsy cores (%), median (IQR)	21.3 (7.1–36.7)
Presence of upgrading at final pathology (n, %)	
Yes	75 (47.2)
No	84 (52.8)

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: the Prostate Imaging Reporting and Data System.

Table 4
Global grade groups on biopsy and radical prostatectomy and change in grade

		GS (GG) at RP (N [% of GS/GG])					Change in score (N [% of GS/GG])		
		6	3 + 4	4 + 3	8	9–10			
Biopsy GS (GG)	N	(GG1)	(GG2)	(GG3)	(GG4)	(GG5)	Upgrade	No Change	Downgrade
6 (GG1)	166	46 (27.7)	83 (50.0)	21 (12.7)	11 (6.6)	5 (3.0)	120 (72.3)	46 (27.7)	-
3 + 4 (GG2)	138	5 (3.6)	58 (42.0)	43 (31.2)	17 (12.3)	15 (10.9)	75 (54.4)	58 (42.0)	5 (3.6)
4 + 3 (GG3)	111	6 (5.4)	19 (17.1)	53 (47.7)	20 (18.0)	13 (11.8)	33 (29.8)	53 (47.7)	25 (22.5)
8 (GG4)	115	1 (0.9)	8 (7.0)	21 (18.3)	51 (44.3)	34 (29.5)	34 (29.5)	51 (44.3)	30 (26.2)
Total	530	58 (10.9)	168 (31.7)	138 (26.0)	99 (18.7)	67 (12.7)	262 (49.4)	208 (39.2)	60 (11.3)

GS: Gleason score; GG: Gleason grade group.

Ml-assisted Models

In multivariable analysis, %fPSA (>0.16 versus ≤0.16) (OR 0.52; 95%CI: 0.27–0.995; $P = 0.048$), apical involvement (No versus Yes) (OR 1.80; 95%CI: 1.02–3.19; $P = 0.042$) on MRI and biopsy GG 1 ($P < 0.001$) were significantly associated with upgrading at RP (Table 5). According to their respective coefficients, the LR model was constructed using the following formula: $Y = 2.29 - 0.65 \times \log(\%fPSA) - 0.59 \times \log(\text{apical involvement}) - 0.97 \times \log(\text{biopsy GG})$.

Table 5
Factors associated with upgrading on univariable and multivariable logistic regression analyses

	Univariable analysis		Multivariable analysis	
	OR (95%CI)	*p* value	OR (95%CI)	*p* value
Age (y)	0.97 (0.94–0.99)	0.011	0.97 (0.94–1.001)	0.055
PV (ml)	0.99 (0.98–1.01)	0.287	1.004 (0.99–1.02)	0.527
D-max (cm)	0.95 (0.77–1.18)	0.659	0.92 (0.64–1.30)	0.628
TPSA (ng/ml)	1.002 (0.997–1.01)	0.358	1.00 (0.99–1.01)	0.892
%fPSA				
≤ 0.16	1 (reference)		1 (reference)	
> 0.16	0.58 (0.35–0.96)	0.035	0.52 (0.27–0.995)	0.048
PSAD				
≤ 0.20	1 (reference)		1 (reference)	
> 0.20	1.42 (0.77–2.60)	0.258	1.44 (0.63–3.29)	0.389
PI-RADS score				
< 3	1 (reference)		1 (reference)	
3	1.11 (0.38–3.22)	0.846	1.87 (0.53–6.53)	0.329
4	1.88 (0.70–5.03)	0.212	2.55 (0.76–8.50)	0.129
5	1.56 (0.61–3.96)	0.353	2.40 (0.69–8.30)	0.167
Apical involvement				
Yes	1 (reference)		1 (reference)	
No	1.21 (0.80–1.82)	0.365	1.80 (1.02–3.19)	0.042
Clinical T stage at MRI				
T1-2	1 (reference)		1 (reference)	
T3a	1.86 (0.99–3.48)	0.054	2.53 (1.07–5.99)	0.034
T3b	1.28 (0.80–2.06)	0.304	1.67 (0.73–3.79)	0.223

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: Prostate Imaging Reporting and Data System.

	Univariable analysis		Multivariable analysis	
T4	0.57 (0.05–6.36)	0.647	4.41 (0.30-64.07)	0.277
No. of positive cores	1.07 (1.01–1.13)	0.025	1.03 (0.88–1.21)	0.676
Presence of csPCa at core				
Yes	1 (reference)		1 (reference)	
No	2.02 (1.18–3.45)	0.011	0.57 (0.20–1.61)	0.285
Presence of core with tumor length > 0.6 cm				
Yes	1 (reference)		1 (reference)	
No	0.95 (0.58–1.55)	0.835	0.97 (0.40–2.34)	0.944
Maximum tumor length in single core	1.22 (0.82–1.82)	0.326	0.74 (0.29–1.91)	0.535
Total tumor length	1.08 (1.02–1.15)	0.014	1.12 (0.88–1.44)	0.364
Percentage of tumor in total biopsy cores	1.01 (1.003–1.03)	0.014	1.01 (0.97–1.06)	0.557
Biopsy grade group				
1	1 (reference)		1 (reference)	
2	0.52 (0.29–0.92)	0.024	0.17 (0.06–0.44)	< 0.001
3	0.13 (0.07–0.26)	< 0.001	0.04 (0.02–0.12)	< 0.001
4	0.16 (0.09–0.30)	< 0.001	0.05 (0.02–0.13)	< 0.001

PV: prostate volume; D-max: maximum diameter of the index lesion on MRI; PSA: prostate-specific antigen; PSAD: prostate-specific antigen density; csPCa: clinically significantly prostate cancer; PI-RADS: the Prostate Imaging Reporting and Data System.

Based on the results of Lasso regression analysis, those clinical features with coefficients >0.1 were selected as the parameters included in the construction of Lasso-LR model. Finally, %fPSA, apical involvement, PI-RADS score, clinical T stage and biopsy GG were the selected features (Fig. 1). The Lasso-LR was constructed by using the following formula: $Y = 1.81 - 0.41 \times \log(\%fPSA) - 0.25 \times \log(\text{apical involvement}) - 0.81 \times \log(\text{biopsy GG}) + 0.15 \times \log(\text{clinical T stage}) + 0.11 \times \log(\text{PI-RADS score})$.

The process of feature selection by RF model and the importance of features are illustrated in Fig. 2. Based on different combinations of clinical parameters, each tree in the forest votes for the major classification, and the final classification of the RF model is derived from the majority of these votes [Loading [MathJax]/jax/output/CommonHTML/jax.js] best number of variables tried at each split were 131 and 4,

respectively. The out of bag (OOB) estimate of error rate was 33.42%, suggesting that the generalization error was quite unsatisfactory.

In the RFE-SVM analysis, 10 clinical parameters were selected as the final candidates for constructing the predictive model without impacting the prediction accuracy of the model, including biopsy GG, apical involvement, maximum tumor length in single core, %fPSA, PSAD, presence of core with tumor length >0.6 cm, presence of csPCa at core, PI-RADS score, D-max and clinical T stage (Fig. 3a). As depicted in Fig. 3b, with the selected features being added to the SVM model one by one, the AUC value of model also increased little by little.

Comparison Between Ml-based Models

Among these models, Lasso-LR model had the highest AUC (0.776, 95%CI: 0.729–0.822), followed by SVM (AUC 0.740, 95%CI: 0.690–0.790), LR (AUC 0.725, 95%CI: 0.674–0.776) and RF (AUC 0.666, 95%CI: 0.618–0.714) (Fig. 4a). Similarly, in the testing dataset, Lasso-LR model had the highest AUC (0.735, 95%CI: 0.656–0.813), followed by SVM (AUC 0.723, 95%CI: 0.644–0.802), LR (AUC 0.697, 95%CI: 0.615–0.778) and RF (AUC 0.607, 95%CI: 0.531–0.684) (Fig. 4b). The Lasso-LR model illustrated an accuracy of 0.712, a sensitivity of 0.679 and a specificity of 0.745, indicating that this model correctly identified 67.9% of PCa patients who experienced upgrading at RP and 74.5% of PCa patients who did not experience upgrading at RP (Table 6). In addition, the Lasso-LR model had the highest YI (0.424) compared with other models. Due to the fact that the YI was calculated as a summation of the sensitivity and specificity minus 1, the highest YI indicated that both the sensitivity and specificity of the Lasso-LR model are reasonably good relative to other models. Pairwise comparison of ROC curves showed that the AUC of Lasso-LR model was significantly higher than that of LR ($P = 0.002$), while the AUCs of SVM and RF were not significantly different to that of LR ($P > 0.05$) (Fig. 4a).

Table 6
Discrimination of prediction models

	LR	Lasso-LR	SVM	RF
Accuracy	0.679	0.712	0.701	0.666
Sensitivity	0.663	0.679	0.668	0.679
Specificity	0.696	0.745	0.734	0.652
YI	0.359	0.424	0.402	0.331
PPV	0.689	0.730	0.718	0.665
NPV	0.670	0.695	0.685	0.667
AUC	0.725	0.776	0.740	0.666

YI: Youden index; PPV: positive predictive value; NPV: negative predictive value; AUC: area under the ROC curve.

Figure 2. Results of feature selection, feature ranking, and model construction with RFE-SVM analysis. (**a**) Distribution of weight for features with RFE-SVM analysis. (**b**) RFE-SVM classifier is trained by adding ranked feature one by one. The iteration repeated until the desired number of features was reached.

Figure 3. Results of model analysis with RFs. (**a**) The detail distribution of classification trees. (**b**) The importance of features ranked by mean decrease accuracy and mean decrease Gini.

Figure 4. The ROC results of ML-based models in the training dataset (**a**) and testing dataset (**b**).

Figure 5. Calibration plots of LR (**a**), Lasso-LR (**b)** SVM (**c**) and RF (**d**).

The calibration of ML-based models was evaluated graphically by the formulation of calibration curves (Fig. 5). The green line represented the fit of the model. Deviations from the 45° line indicated miscalibration. Part of the green line below the 45° line indicated that higher predicted probabilities might overestimate the true outcome, and part of green line upon the 45° line indicated that lower predicted probabilities might under-predict the true probability of upgrading. The Lasso-LR model was well-calibrated (Fig. 5b), followed by SVM (Fig. 5d), RF (Fig. 5c) and LR (Fig. 5a).

Discussion

Accurate risk stratification, which mainly depends on PSA level, biopsy grade group and stage classification, plays a pivotal role in guiding treatment management for PCa patients. However, it has been demonstrated that prostate biopsy often underestimates the cancer and incorrectly assigns NCCN risk stratification [3, 4]. There are several reasons accounting for the discrepancies between the biopsy and RP grades: sampled cores and indications for biopsy, differences in biopsy techniques, erroneous diagnostic interpretation, tumor heterogeneity, sampling error on biopsy, clinician interpretation of the biopsy GG

ore) and practice variations regarding biopsy grade

assignment [7, 23]. The incorrect risk stratification may impact treatment planning, patient selection and decision-making processes. Therefore, it is extremely important to the identify risk factors associated with upgrading to avoid under-treatment, especially among those PCa patients who are considered appropriate candidates for AS. Unfortunately, there are currently no widely accepted predictive models to accurately predict the final individualized GG at RP and the discrimination ability of various models remains modest [24–27]. Machine learning has been previously used for predicting outcomes in other fields of medicine, including the identification of lung cancer based on routine blood indices and the in-hospital rupture of type A aortic dissection [28, 29]. Given the excellent performance of machine learning algorithms in classification, four machine learning algorithms were employed in our study to determine relevant risk factors; we then developed and validated four novel prediction models to identify those PCa patients at high risk of harboring upgrading at RP before making treatment decisions.

Overall, up to 49.4% included patients were upgraded at RP, especially biopsy GG 1 patients, with the proportion of upgrading being 72.3%. Similarly, Altok and colleagues [30] reported that 70.9% of biopsy GG 1 patients in their study cohort were upgraded at RP, and most were upgraded to GG 2. These observations explain why some patients with GG 1 disease at biopsy suffer metastases or die of prostate cancer and suggest that a substantial proportion of biopsy GG 1 patients who embark on active surveillance are not, in fact, suitable candidates [31]. Given the known risk of underestimation in biopsy specimens, the prediction of GG upgrade plays a major role when considering individualized therapy for PCa patients, especially AS [32]. In our series, %fPSA (>0.16 versus ≤0.16), apical involvement at MRI (No versus Yes) and biopsy grade group (GG 4, GG 3, GG 2 versus GG 1) were independent factors in multivariable logistic regression analysis. %fPSA, apical involvement at MRI, biopsy grade group and clinical T stage at MRI were significantly associated with upgrading in Lasso-LR and SVM model. However, in a comparable study, Alshak et al. [22] demonstrated that only the PI-RADS score was a significant predictor of upgrading. Besides, Gandaglia et al. [33] reported that preoperative PSA level, GG at MRI-targeted biopsy and clinically significant PCa at systematic biopsy were independent risk factors of upgrading at RP. The differences in results between our study and the latter two studies might be due to the fact that the latter two studies did not include detailed core biopsy information, which has been successfully shown to contain huge potential predictive value.

In our study, imaging factors such as apical involvement at MRI and clinical T stage at MRI, were more important predictors than clinical parameters according to the results of ML-based feature ranking analyses, except for RF analysis. This implied that mp-MRI had great potential in predicting upgrading, irrespective of its important role in detecting csPCa and assigning accurate risk stratification for PCa patients. The routine mp-MRI examination for patients with suspected PCa before biopsy was indeed beneficial and helpful. Among those biopsy-related variables, biopsy GG was always the strongest predictor. In the LR and Lasso-LR model, the number of positive cores, presence of csPCa at core, presence of a core with a tumor length >0.6 cm, maximum tumor length in a single core, total tumor length and percentage of tumor in total biopsy cores demonstrated almost no value in the prediction of upgrading at RP, while the number of positive cores, total tumor length and percentage of tumor in total biopsy cores [illegible] features, including D-max, were reliable proxies of tumor volume,

the size of tumor should not be considered relevant to the presence of upgrading [34]. Nonetheless, Corcoran et al. [23] reported that tumor volume of PCa was a significant predictor of upgrading in multivariable analysis, and the measurement of surrogate of tumor volume might predict those at greatest risk of Gleason score upgrade. One thing to be noted was that the patient cohort in the study of Corcoran et al. [23] did not include those patients with biopsy GG 3 and 4. %fPSA outperformed TPSA and PSAD in the prediction of upgrading in the LR, Lasso-LR and SVM models. On the contrary, in the mean decrease accuracy and mean decrease Gini evaluation of RF models, TPSA and PSAD ranked higher than %fPSA.

For the performance of ML-based models, the Lasso-LR model showed the best discriminative power with an AUC of 0.776 (95%CI: 0.729–0.822), followed by SVM (AUC 0.740; 95%CI: 0.690–0.790), LR (AUC 0.725, 95%CI: 0.674–0.776) and RF (AUC 0.666; 95%CI: 0.618–0.714). The nomogram developed by He et al. [35] achieved an AUC of 0.753 in the prediction of upgrading, which was higher than that of LR but lower than Lasso-LR in the present study. Also, Moussa et al. [36] constructed a normogram for predicting the possibility of upgrading, with a concordance index of 0.68. Additionally, all of the ML-based models except for RF outperformed the predictive models constructed by Kulkarni et al. [37] and Athanazio et al. [7], with AUC values of 0.71 and 0.699 in the respective studies. Of note, in a study consisting of 2982 PCa patients treated with RP, the model for predicting upgrading based on logistic regression analysis showed a predictive accuracy of 0.804; in contrast, in our study, the Lasso-LR model presented the best predictive accuracy of 0.712 [26]. Despite the better predictive accuracy of the model in the study of Chun et al. [26], it was still difficult to determine the model with best performance when compared with our ML-based models as there was no other discrimination metrics such as AUC, sensitivity, specificity, PPV and NPV in their study. It should be noted that the good performance of our ML-based models might be related to the inclusion of mp-MRI information and detailed biopsy information.

Despite several strengths, our study has certain limitations. First, the data on PCa patients who underwent RP enrolled in our study cohort were retrospectively collected at a single institution, which may have resulted in selection bias. Second, the case-level highest Gleason grade group was more commonly assigned to patients undergoing systematic TRUS-guided biopsy in our country; hence, we should also construct predictive models to identify risk factors associated with upgrading using a comparison between the highest biopsy GG and final RP samples.

Conclusions

In summary, we developed four ML-based models to help clinicians identify the individualized risk of upgrading for PCa patients after prostate needle biopsy. The Lasso-LR model had the best discriminative power according to the results of pairwise comparisons. We believe that our research findings can make a significant difference in the process of treatment decision making by more accurately identifying patients at high risk of harboring upgrading at RP. Of course, further validation in multiple institutions with a large sample size is warranted.

Abbreviations

AS
Active surveillance; AUC:Area under the ROC curve; csPCa:Clinically significantly prostate cancer; D-max:Maximum diameter of the index lesion on MRI; GG:Gleason grade group; Lasso:least absolute shrinkage and selection operator; LR:Logistic regression; ML-Machine learning; mp-MRI:Multi-parametric MRI; NPV:Negative predictive value; PI-RADS:The Prostate Imaging Reporting and Data System; PPV:Positive predictive value; PSA:Prostate-specific antigen; PSAD:Prostate-specific antigen density; PV:Prostate volume; RF:Random forest; SVM:Super vector machine; YI:Youden index.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

HLL and ZQC designed the research. KT, EJP and LW collected, analyzed and interpreted the clinical data. HLL and DX contributed to the drafting of the manuscript. ZQC and DX revised the manuscript. Both ZQC and DX are the corresponding authors. All authors approved the final manuscript.

Funding

No funding.

Availability of data and materials

The data used to support the findings of this study are available from the corresponding author upon request.

Ethics approval and consent to participate

The present study was approved by the institutional review board of Tongji Hospital. This study does not involve the use of any animal data or tissue.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974;111:58-64.
2. Epstein JI, Egevad L, Amin, MB, Delahunt B, Srigley JR, Humphrey PA. 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;40:244-52.
3. Alchin DR, Murphy D, Lawrentschuk N. Risk factors for Gleason Score upgrading following radical prostatectomy. Minerva Urol Nefrol 2017, 69:459-465.
4. Müntener M, Epstein JI, Hernandez DJ, Gonzalgo ML, Mangold L, Humphreys E, Walsh PC, Partin AW, Nielsen ME. Prognostic significance of Gleason score discrepancies between needle biopsy and radical prostatectomy. Eur Urol 2008, 53:767-775.
5. Hamdy FC, Donovan JL, Lane JA, Mason M, Metcalfe C, Holding P, Davis M, Peters TJ, Turner EL, Martin RM, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. N Engl J Med 2016, 375:1415-1424.
6. Qi F, Zhu K, Cheng Y, Hua L, Cheng G. How to Pick Out the "Unreal" Gleason 3 + 3 Patients: A Nomogram for More Precise Active Surveillance Protocol in Low-Risk Prostate Cancer in a Chinese Population. J Invest Surg 2019:1-8.
7. Athanazio D, Gotto G, Shea-Budgell M, Yilmaz A, Trpkov K. Global Gleason grade groups in prostate cancer: concordance of biopsy and radical prostatectomy grades and predictors of upgrade and downgrade. Histopathology 2017, 70:1098-1106.
8. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med 2018, 24:1304-1305.
9. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017, 376:2507-2509.
10. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019, 380:1347-1358.
11. Trpkov K, Sangkhamanon S, Yilmaz A, Medlicott SAC, Donnelly B, Gotto G, Shea-Budgell M. Concordance of "Case Level" Global, Highest, and Largest Volume Cancer Grade Group on Needle Biopsy Versus Grade Group on Radical Prostatectomy. Am J Surg Pathol 2018, 42:1522-1529.
12. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2016, 69:16-40.
13. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, Vickers AJ, Parwani AV, Reuter VE, Fine SW, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur Urol 2016, 69:428-435.
14. Hou Y, Bao ML, Wu CJ, Zhang J, Zhang YD, Shi HB. A machine learning-assisted decision-support model to better identify patients with prostate cancer requiring an extended pelvic lymph node dissection. BJU Int 2019, 124:972-983.

15. Archer KJ, Williams AA. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. Stat Med 2012, 31:1464-1474.
16. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res 2020:962280220921415.
17. Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. Bioinformatics 2015, 31:i197-205.
18. Wong NC, Lam C, Patterson L, Shayegan B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. BJU Int 2019, 123:51-57.
19. Van Belle V, Van Calster B, Van Huffel S, Suykens JA, Lisboa P. Explaining Support Vector Machines: A Color Based Nomogram. PLoS One 2016, 11:e0164568.
20. Guyon I, Wston J, Barnhill S. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002, 46:p.389-422.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988, 44:837-845.
22. Alshak MN, Patel N, Gross MD, Margolis D, Hu JC. Persistent Discordance in Grade, Stage, and NCCN Risk Stratification in Men Undergoing Targeted Biopsy and Radical Prostatectomy. Urology 2020, 135:117-123.
23. Corcoran NM, Hovens CM, Hong MK, Pedersen J, Casey RG, Connolly S, Peters J, Harewood L, Gleave ME, Goldenberg SL, Costello AJ. Underestimation of Gleason score at prostate biopsy reflects sampling error in lower volume tumours. BJU Int 2012, 109:660-664.
24. Kuroiwa K, Shiraishi T, Naito S. Gleason score correlation between biopsy and prostatectomy specimens and prediction of high-grade Gleason patterns: significance of central pathologic review. Urology 2011, 77:407-411.
25. Epstein JI, Feng Z, Trock BJ, Pierorazio PM. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. Eur Urol 2012, 61:1019-1024.
26. Chun FK, Steuber T, Erbersdobler A, Currlin E, Walz J, Schlomm T, Haese A, Heinzer H, McCormack M, Huland H, et al. Development and internal validation of a nomogram predicting the probability of prostate cancer Gleason sum upgrading between biopsy and radical prostatectomy pathology. Eur Urol 2006, 49:820-826.
27. Thomas C, Pfirrmann K, Pieles F, Bogumil A, Gillitzer R, Wiesner C, Thüroff JW, Melchior SW. Predictors for clinically relevant Gleason score upgrade in patients undergoing radical prostatectomy. BJU Int 2012, 109:214-219.
28. Wu J, Qiu J, Xie E, Jiang W, Zhao R, Qiu J, Zafar MA, Huang Y, Yu C. Predicting in-hospital rupture of type A aortic dissection using Random Forest. J Thorac Dis 2019, 11:4634-4646.
29. Wu J, Zan X, Gao L, Zhao J, Fan J, Shi H, Wan Y, Yu E, Li S, Xie X. A Machine Learning Method for [illegible] ine Blood Indices: Qualitative Feasibility Study. JMIR Med

Inform 2019, 7:e13476.

30. Altok M, Troncoso P, Achim MF, Matin SF, Gonzalez GN, Davis JW. Prostate cancer upgrading or downgrading of biopsy Gleason scores at radical prostatectomy: prediction of "regression to the mean" using routine clinical features with correlating biochemical relapse rates. Asian J Androl 2019, 21:598-604.
31. Yang DD, Mahal BA, Muralidhar V, Vastola ME, Boldbaatar N, Labe SA, Nezolosky MD, Orio PF, 3rd, King MT, Martin NE, et al. Pathologic Outcomes of Gleason 6 Favorable Intermediate-Risk Prostate Cancer Treated With Radical Prostatectomy: Implications for Active Surveillance. Clin Genitourin Cancer 2018, 16:226-234.
32. Morlacco A, Cheville JC, Rangel LJ, Gearman DJ, Karnes RJ. Adverse Disease Features in Gleason Score 3 + 4 "Favorable Intermediate-Risk" Prostate Cancer: Implications for Active Surveillance. Eur Urol 2017, 72:442-447.
33. Gandaglia G, Ploussard G, Valerio M, Mattei A, Fiori C, Roumiguie M, Fossati N, Stabile A, Beauval JB, Malavaud B, et al. The Key Combined Value of Multiparametric Magnetic Resonance Imaging, and Magnetic Resonance Imaging-targeted and Concomitant Systematic Biopsies for the Prediction of Adverse Pathological Features in Prostate Cancer Patients Undergoing Radical Prostatectomy. Eur Urol 2020, 77:733-741.
34. Gandaglia G, Ploussard G, Valerio M, Mattei A, Fiori C, Fossati N, Stabile A, Beauval JB, Malavaud B, Roumiguié M, et al. A Novel Nomogram to Identify Candidates for Extended Pelvic Lymph Node Dissection Among Patients with Clinically Localized Prostate Cancer Diagnosed with Magnetic Resonance Imaging-targeted and Systematic Biopsies. Eur Urol 2019, 75:506-514.
35. He B, Chen R, Gao X, Ren S, Yang B, Hou J, Wang L, Yang Q, Zhou T, Zhao L, et al. Nomograms for predicting Gleason upgrading in a contemporary Chinese cohort receiving radical prostatectomy after extended prostate biopsy: development and internal validation. Oncotarget 2016, 7:17275-17285.
36. Moussa AS, Kattan MW, Berglund R, Yu C, Fareed K, Jones JS. A nomogram for predicting upgrading in patients with low- and intermediate-grade prostate cancer in the era of extended prostate sampling. BJU Int 2010, 105:352-358.
37. Kulkarni GS, Lockwood G, Evans A, Toi A, Trachtenberg J, Jewett MA, Finelli A, Fleshner NE. Clinical predictors of Gleason score upgrading: implications for patients considering watchful waiting, active surveillance, or brachytherapy. Cancer 2007, 109:2432-2438.

Figures

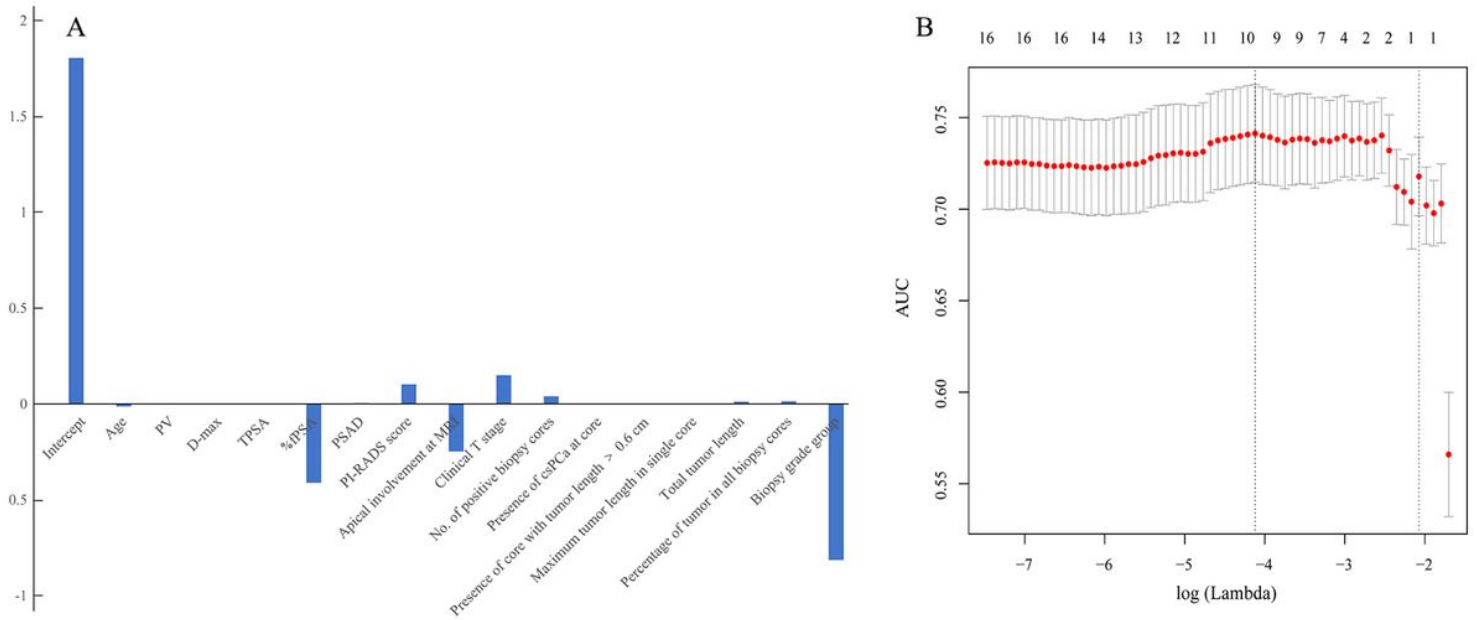

Figure 1

Distribution of feature coefficients estimated by Lasso-LR analysis (a) and the optimal features are those with a coefficient > 0.1 and produced best accuracy (b).

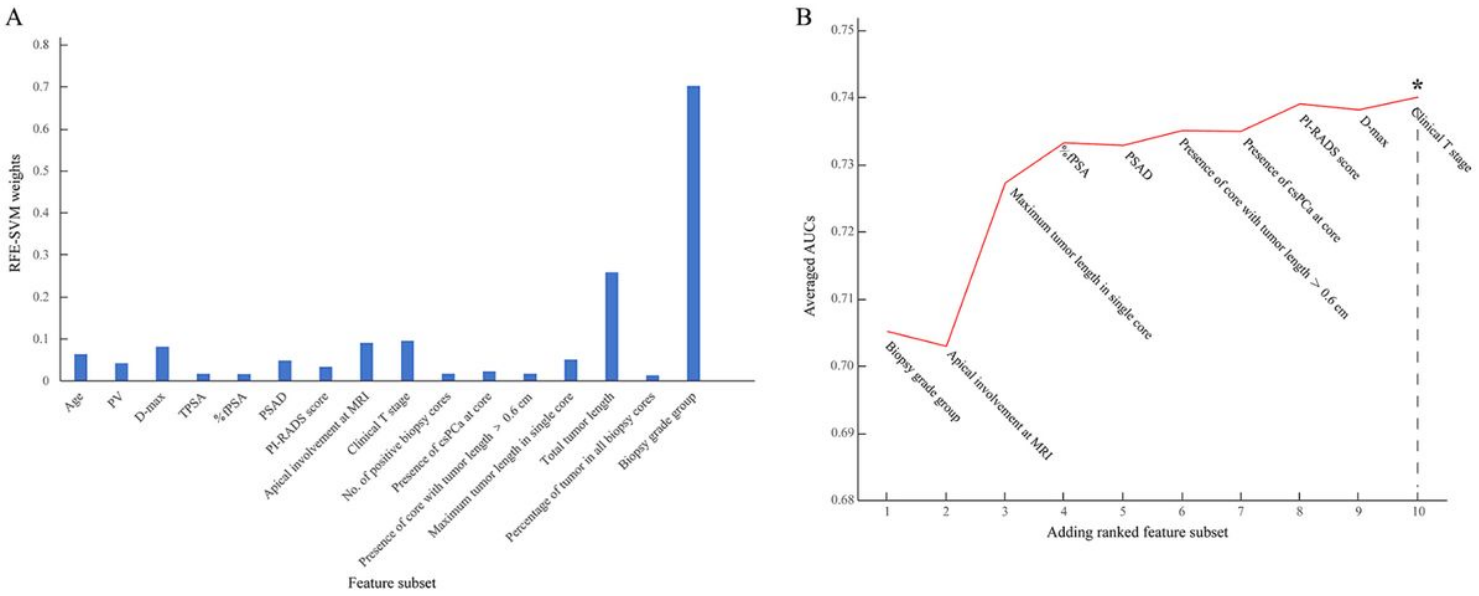

Figure 2

Results of feature selection, feature ranking, and model construction with RFE-SVM analysis. (a) Distribution of weight for features with RFE-SVM analysis. (b) RFE-SVM classifier is trained by adding ranked feature one by one. The iteration repeated until the desired number of features was reached.

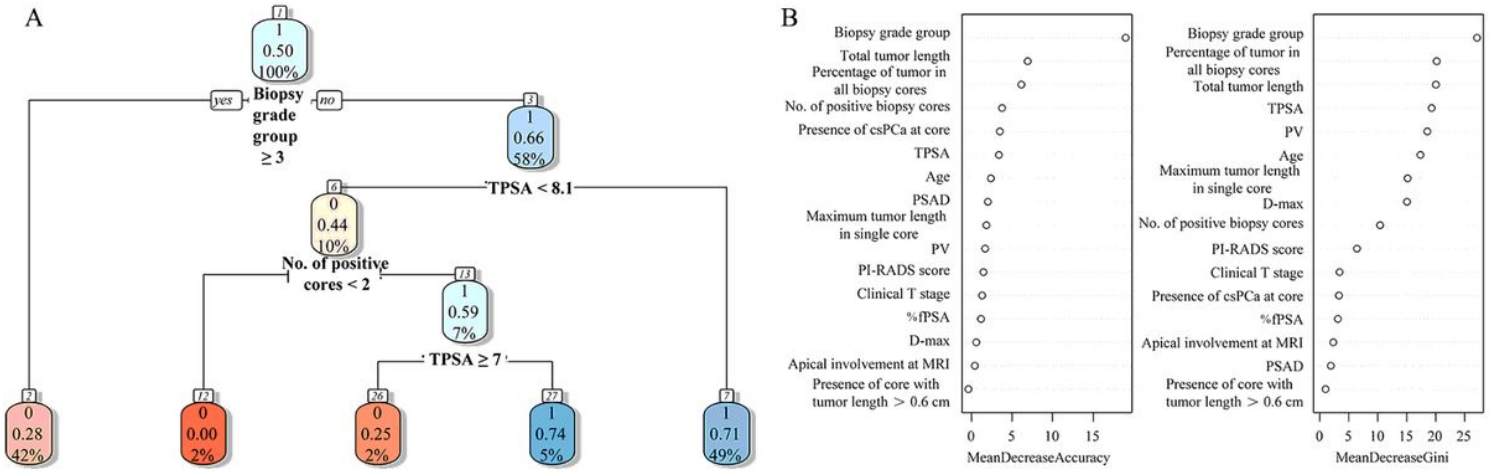

Figure 3

Results of model analysis with RFs. (a) The detail distribution of classification trees. (b) The importance of features ranked by mean decrease accuracy and mean decrease Gini.

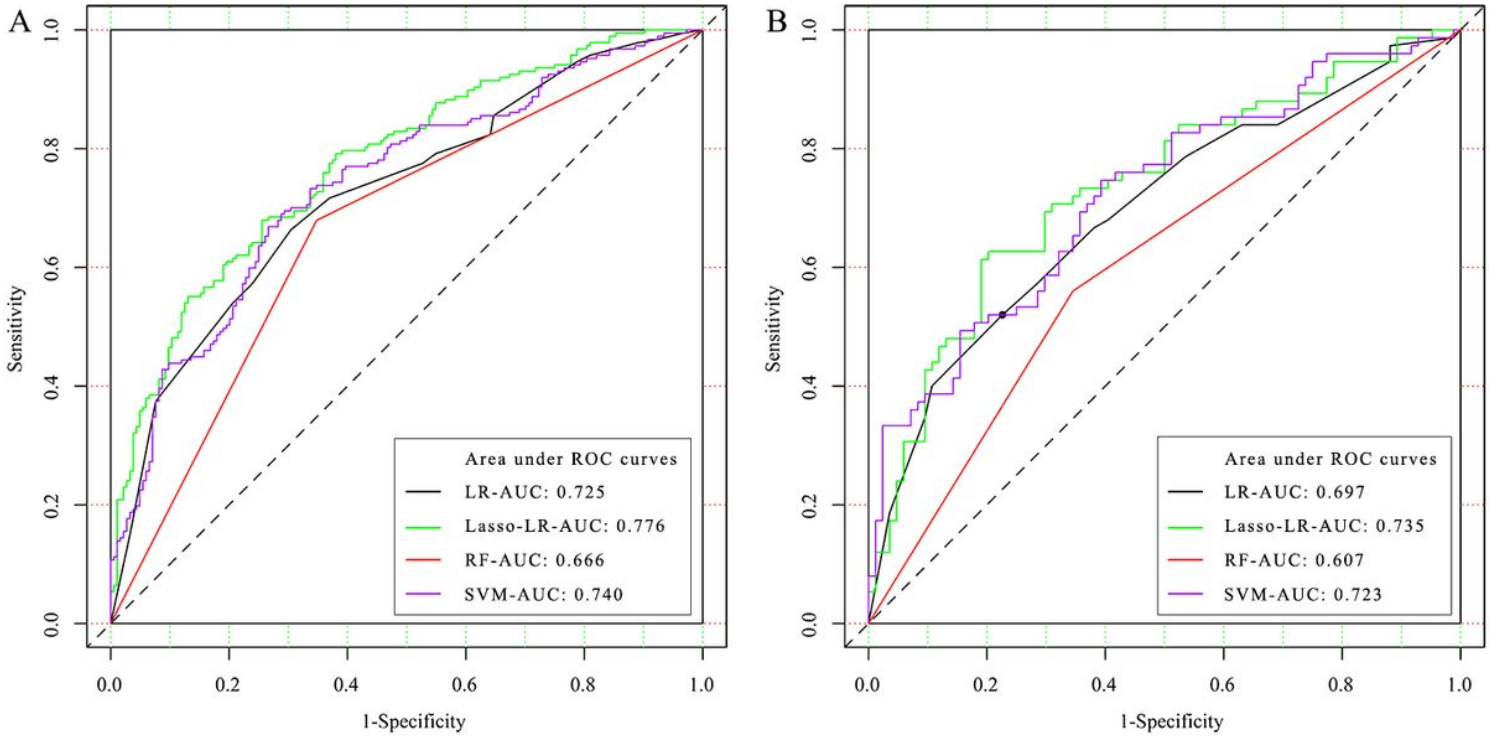

Figure 4

The ROC results of ML-based models in the training dataset (a) and testing dataset (b).

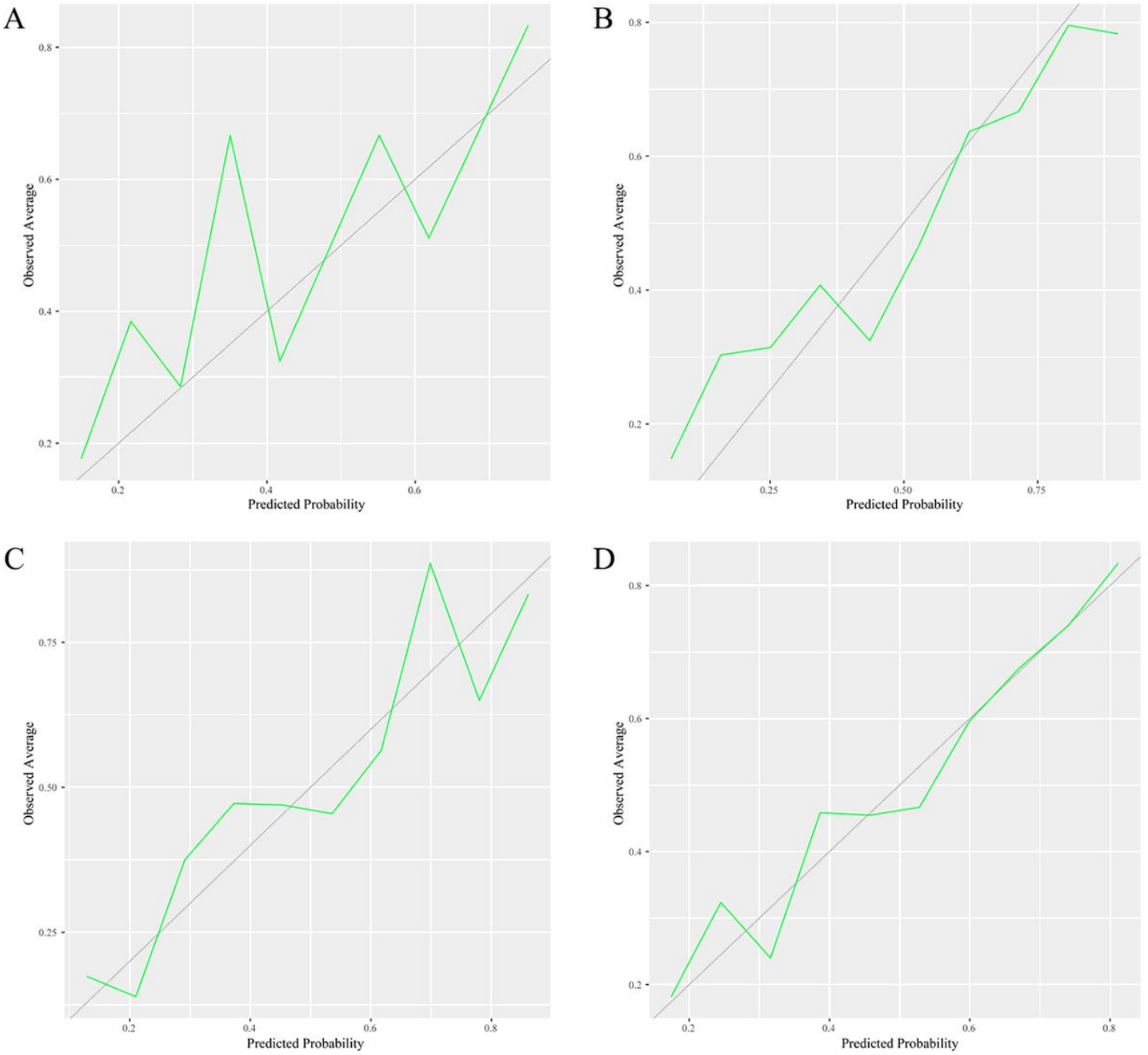

Figure 5

Calibration plots of LR (a), Lasso-LR (b) SVM (c) and RF (d).