Article

Discovery Medicine 2026; 38(214): 3657–3669
https://doi.org/10.24976/Discov.Med.202638214.278

Machine Learning Models for Predicting Upgrading to ISUP Grade Group ≥3 After Radical Prostatectomy in Patients With D'Amico Low-Risk Prostate Cancer

Ismaıl Onder Yılmaz[1], Ozgür Yılmaz[2], Mehmet Zubaroglu[1,3,*], Rahmi Kavak[4], Nebıl Akdogan[1], Mehmet Gurkan Arıkan[1], Mutlu Deger[1], Volkan Izol[1,5]

[1]Department of Urology, Faculty of Medicine, Çukurova University, 01330 Adana, Turkey
[2]Department of Artificial Intelligence Engineering, Adana Alparslan Turkes Science and Technology University, 01250 Adana, Turkey
[3]Department of Urology, Tarsus State Hospital, 33460 Mersin, Turkey
[4]Department of Software Engineering, Adana Alparslan Turkes Science and Technology University, 01250 Adana, Turkey
[5]Private Clinic, Cukurova Urology Center, 01170 Adana, Turkey
*Correspondence: mehmet.zubaroglu@hotmail.com (Mehmet Zubaroglu)
Submitted: 2 July 2026 Revised: 28 August 2026 Accepted: 8 September 2026 Published: 25 November 2026

Background: Patients initially classified as having low-risk prostate cancer may be found to have higher-grade disease at radical prostatectomy, highlighting the need for improved preoperative risk stratification. This study aimed to develop and internally evaluate machine learning models using preoperative variables to predict major upgrading to International Society of Urological Pathology (ISUP) Grade Group ≥3 after radical prostatectomy among surgically treated patients with D'Amico low-risk prostate cancer.

Methods: This retrospective single-centre study included 133 surgically treated patients with D'Amico low-risk prostate cancer who underwent radical prostatectomy. Major upgrading was defined as postoperative ISUP Grade Group ≥3 and occurred in 13 patients (9.8%). Sixteen demographic, clinical, laboratory, and biopsy-derived variables were evaluated using 11 supervised machine learning algorithms. Repeated stratified 5-fold cross-validation with 10 repetitions was used for primary internal validation. Hyperparameter optimisation, when performed, was embedded within nested cross-validation, and bootstrap resampling was used as a complementary robustness analysis. Performance was assessed using the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision–recall curve (PR-AUC), and the Brier score.

Results: There were significant differences in body mass index, prostate volume, clinical T-stage distribution, and minimum tumour percentage per core. Random Forest yielded the highest mean ROC-AUC (0.83 ± 0.09), followed by Gaussian Naïve Bayes (0.78 ± 0.10). Logistic Regression yielded a mean ROC-AUC of 0.74 ± 0.11; the numerical difference of 0.09 was not interpreted as evidence of model superiority. Exploratory feature importance analysis ranked prostate volume, biopsy-core measures, body mass index, prostate-specific antigen, and minimum tumour percentage per core among the higher-ranked variables.

Conclusion: Machine learning may help integrate routinely available preoperative variables for predicting major upgrading, but these exploratory findings require external validation in larger multicentre cohorts.

Keywords: machine learning; ISUP grade upgrading; radical prostatectomy; prostate cancer risk stratification; random forest algorithm

Introduction

Prostate cancer is one of the most common malignancies in men, with approximately 1.4 million new cases and 375,000 deaths reported worldwide in 2020 [1]. In Europe, it is the most frequently diagnosed cancer in men and the third leading cause of cancer-related mortality [2].

The D'Amico classification is a widely used risk assessment system in prostate cancer treatment, categorising patients into low-, intermediate-, and high-risk groups based on prostate-specific antigen (PSA) levels, Gleason scores, and clinical stages. For the present study, low-risk prostate cancer was defined using a D'Amico-based risk framework, requiring a PSA level ≤10 ng/mL, a Gleason score ≤6 (International Society of Urological Pathology [ISUP] Grade Group 1), and clinical stage T1–T2a [3,4].

The Gleason score is a histological grading measure based on the architectural patterns of prostatic adenocarcinoma. The ISUP Grade Group system comprises five groups derived from Gleason scores, with a Gleason score ≤6 corresponding to Grade Group 1 [5]. Approximately 29% of prostate cancer diagnoses fall into the D'Amico low-risk group [6]. The European Association of Urology (EAU) Guidelines base prostate cancer diagnosis and treatment strategies on the D'Amico classification [3]. Additionally, the risk classifications in the National Compre-

Copyright: © 2026 The Author(s). Published by Discovery Medicine. This is an open access article under the CC BY 4.0 license.
Note: *Discovery Medicine* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

hensive Cancer Network (NCCN) and American Urological Association (AUA) Guidelines closely resemble the D'Amico classification [7].

According to the EAU Prostate Cancer Guidelines, radical prostatectomy is considered an effective treatment option for all risk groups in localised prostate cancer [3]. Active surveillance is now the preferred approach for low-risk patients; however, radical prostatectomy remains a valid option that provides durable tumour control and may reduce cancer-specific mortality. Nonetheless, it can result in overtreatment in some patients [8]. However, patients initially classified as low-risk may be found to have a higher ISUP grade upon postoperative pathological evaluation. This highlights the need for more accurate preoperative risk stratification. Studies have reported a Gleason score upgrade rate of 32.2% in patients with low-risk prostate cancer following radical prostatectomy [9].

A previous study has shown that pathological stage upgrading, even in low-risk patients, can impact treatment outcomes [10]. Pathological upgrading is linked to a higher risk of biochemical recurrence (a post-treatment increase in PSA) and adverse features such as positive margins and extracapsular extension, suggesting more aggressive disease. While its effect on cancer-specific and overall survival is uncertain, advanced stage is associated with worse long-term outcomes [11].

Analysing the factors predicting ISUP Grade Group ≥3 upgrading after radical prostatectomy in low-risk patients may enable the development of more accurate prognostic models for patient management. Studies conducted for this purpose have identified predictive factors such as the PSA level (>5 ng/mL), the proportion of positive biopsy cores (>0.20), prostate volume, pathological tumour stage (pT), the pathological Gleason score, tumour burden, and patient risk classification [12,13].

Machine learning (ML) and artificial intelligence (AI) algorithms have been widely used in the medical field in recent years for various purposes, including disease diagnosis, prognosis, risk stratification, and treatment selection [14]. In patients with prostate cancer, numerous studies have utilised ML for various purposes, including histological diagnosis and grade determination [15], prediction of lymph node metastasis [16], and diagnosis based on radiomic and genomic features [17]. A study utilising ML to predict ISUP grade upgrading after radical prostatectomy remains limited [18].

The ability to predict which patients classified as low-risk based on biopsy results will experience ISUP grade upgrading after radical prostatectomy may improve the accuracy of clinical decision-making and help prevent unnecessary or inadequate treatment approaches. Conventional logistic regression models generally assume predominantly linear and additive relationships between predictors and outcome. However, in this specific clinical setting, the risk of ISUP Grade Group ≥3 upgrading is likely influenced by complex and non-linear interactions among PSA-related variables, prostate volume, biopsy core burden, and clinical T stage. Therefore, ML methods may provide a more flexible framework for capturing these interactions and exploring whether complex and potentially non-linear relationships can be captured beyond a conventional linear baseline approach. In this study, we aimed to develop and internally evaluate ML models using preoperative variables to predict major upgrading to ISUP Grade Group ≥3 after radical prostatectomy among surgically treated patients with D'Amico low-risk prostate cancer.

Methods

Patient Selection and Management of Missing Data

The inclusion criteria were: (1) radical prostatectomy performed between January 2011 and June 2024; (2) D'Amico low-risk prostate cancer, defined for the present study as a PSA level ≤10 ng/mL, a Gleason score ≤6 (ISUP Grade Group 1) [5], and clinical stage T1–T2a [3,4]; and (3) availability of complete data for the candidate predictors and postoperative pathological outcome. Patients who did not meet the D'Amico low-risk criteria or had missing data in at least one candidate variable were excluded. A total of 632 patients who underwent radical prostatectomy between January 2011 and June 2024 were assessed for eligibility. Of these, 159 patients met the D'Amico low-risk criteria. Among the eligible patients, 26 (16.4%) were excluded from the complete-case analysis because of missing data in at least one candidate variable. Missing data involved prostate volume (n = 8), PSA density (n = 6), the percentage of positive biopsy cores (n = 5), clinical T stage (n = 3), and other variables (n = 4). No imputation of missing data was performed. Consequently, the final complete-case analysis included 133 patients with complete information for all candidate predictors and the study outcome. Accordingly, the target population of the present study was restricted to surgically treated patients with D'Amico low-risk prostate cancer who underwent radical prostatectomy and did not include patients managed with active surveillance or other non-surgical approaches. The patient-selection process is presented in Fig. 1.

The complete-case cohort consisted of 133 patients, including 13 patients with ISUP Grade Group ≥3 upgrading and 120 patients without major upgrading. For the purposes of this study, major upgrading was defined as postoperative ISUP Grade Group ≥3. For historical pathology reports in which Grade Group was not explicitly reported, the corresponding Grade Group was derived from the documented Gleason score and primary and secondary Gleason patterns, as described below. Transition from biopsy ISUP Grade Group 1 to postoperative ISUP Grade Group 2 was not considered the study outcome, because the primary endpoint was deliberately restricted to a more clinically consequential degree of pathological upgrading. Given the ret-

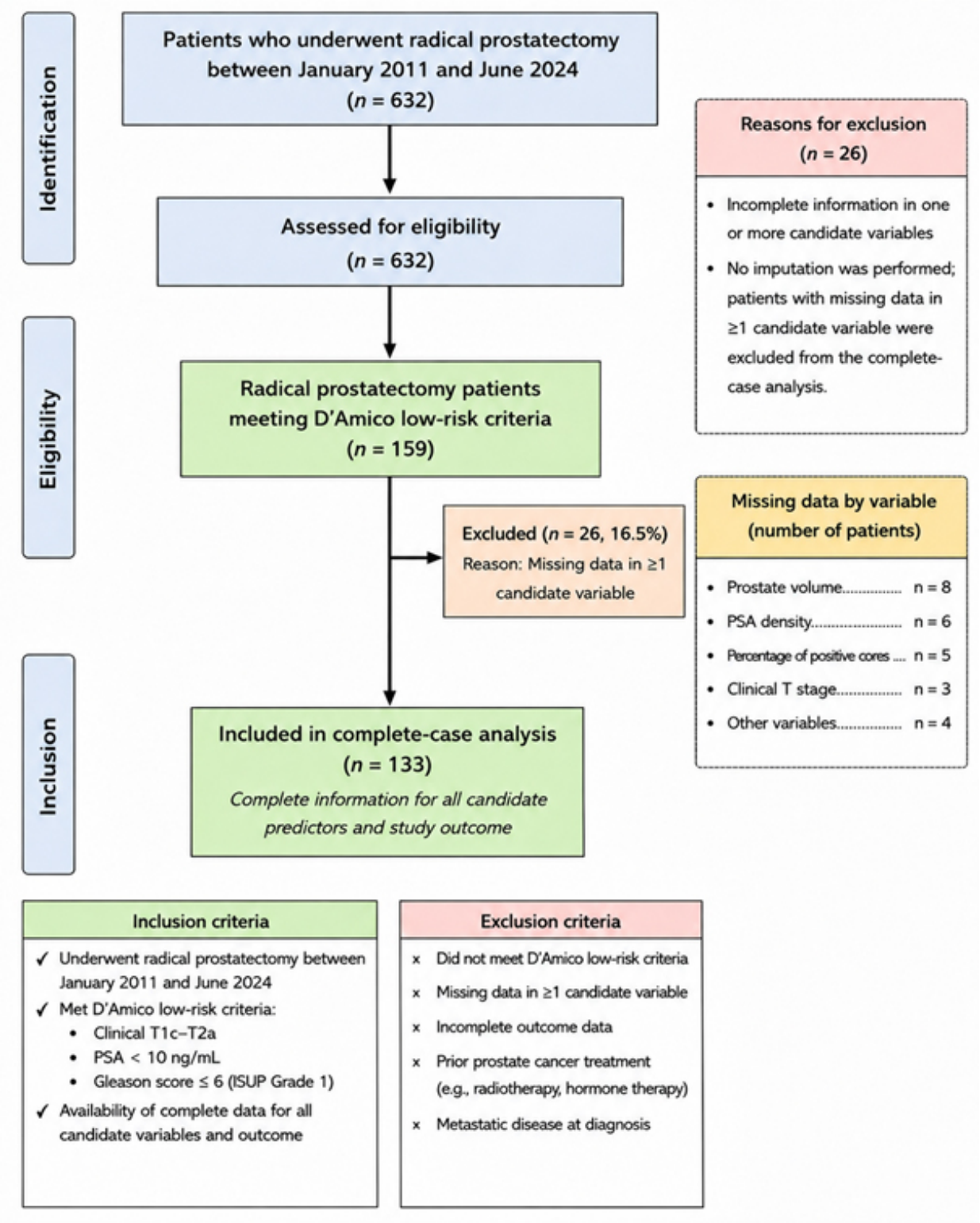


Fig. 1. Flow diagram of patient selection and cohort formation. A total of 632 patients who underwent radical prostatectomy were assessed, 159 met the D'Amico low-risk criteria, and 26 were excluded because of missing data, resulting in a final complete-case cohort of 133 patients.

rospective design and limited number of outcome events, a narrower and less heterogeneous primary endpoint was preferred. All predictive analyses were conducted using the entire cohort of 133 patients within the specified internal validation framework. However, because only 13 outcome events were observed, a conventional multivariable logistic

regression model aimed at estimating independent predictor effects was not fitted, owing to the high risk of unstable coefficient estimates and overfitting. Logistic regression included among the evaluated classifiers was used solely as a predictive baseline model and not for inferential estimation of independent associations. Given the limited number of outcome events, all ML analyses were considered exploratory.

This single-centre retrospective study used data extracted by physicians from the Çukurova University Faculty of Medicine Radical Prostatectomy Database and the hospital automation system. Ethical approval was obtained from the Çukurova University Non-Interventional Clinical Research Ethics Committee (Approval No: 149/68; 8 November 2024). The study complied with the Declaration of Helsinki. The requirement for informed consent was waived by the Çukurova University Non-Interventional Clinical Research Ethics Committee due to the retrospective design, and there was no contact with patients.

Sixteen variables were analysed: demographics (age and body mass index [BMI]); clinical and laboratory measures (prostate volume, PSA, PSA density, and clinical T stage); comorbidities (hypertension and diabetes mellitus); and preoperative biopsy-related features (perineural invasion, the number of biopsy cores, the number and percentage of positive biopsy cores, the maximum and minimum tumour percentage per core, bilateral tumour involvement, and the number of tumour-infiltrated biopsy areas).

ML-Based Classification Methods

Clinical, pathological, and biopsy-related variables were analysed to predict ISUP Grade Group ≥3 upgrading on final pathology using ML. Biopsy and radical prostatectomy pathology results from 2011 to 2024 were obtained from finalised and approved reports in the hospital information system. The specimens were not re-reviewed for the present study.

Because the study period preceded and extended beyond the introduction of the contemporary ISUP Grade Group system, historical pathology reports that did not explicitly state an ISUP Grade Group were harmonised using the recorded Gleason score and, where applicable, the documented primary and secondary Gleason patterns. Gleason score ≤6 (3 + 3) was mapped to Grade Group 1, Gleason score 3 + 4 = 7 to Grade Group 2, Gleason score 4 + 3 = 7 to Grade Group 3, Gleason score 8 to Grade Group 4, and Gleason scores 9–10 to Grade Group 5. This procedure represented retrospective harmonisation of the reported grading categories and not histopathological re-review or retrospective regrading of the original specimens [5]. No attempt was made to reinterpret the original glandular morphology according to later ISUP grading recommendations; the harmonisation was based exclusively on the grading information documented in the original pathology reports.

As part of routine institutional practice, the reports were evaluated and finalised by a single pathologist experienced in uropathology. No study-specific blinding procedure was applied; however, the ML analyses were performed after the reports had been finalised, and the pathologist had no access to the model results. Interobserver agreement was not assessed because no second independent pathological review was performed. The number of tumour-infiltrated biopsy areas was obtained from the preoperative prostate biopsy pathology report and represented the recorded number of biopsy areas containing tumour.

Biopsy technique was recorded for all patients. Based on the hospital records, cognitive transrectal ultrasound (TRUS)-guided biopsy was performed in 87 patients, whereas magnetic resonance (MR) fusion biopsy was performed in 46 patients. The number of biopsy cores was obtained from individual biopsy reports and represented the actual total number of biopsy cores obtained for each patient. The number of biopsy cores varied among patients according to the biopsy technique and clinical practice during the study period. Biopsy technique was not included as a candidate predictor in the original ML feature set. Because biopsy approach also reflected changes in diagnostic practice over the study period and the number of major upgrading events was limited, no post hoc expansion of the candidate feature set was performed by adding biopsy technique. Its potential influence on biopsy-core characteristics and model results was therefore considered as a possible source of residual confounding. Laparoscopic radical prostatectomy was performed in 125 patients and open surgery in 8 patients.

All 16 variables were used as input features, and feature selection was not performed; all raw variables were included in the algorithms. Clinical T stage was entered into the ML models as a single integer-coded feature (T1a = 1, T1b = 2, T1c = 3, and T2a = 4), rather than using dummy/one-hot encoding. This representation imposed a fixed numerical ordering and equal spacing between categories, which should not be interpreted as equivalent differences in clinical risk. No sensitivity analysis using alternative categorical encoding was performed. Binary categorical variables, including hypertension, diabetes mellitus, perineural invasion, and bilateral tumour involvement, were encoded as 0/1 (No = 0, Yes = 1). Continuous variables were standardised using the StandardScaler method from the Scikit-learn library prior to model development. Modern ML approaches such as Random Forest are robust to correlations in feature space; therefore, all variables were retained, and feature importance was evaluated using permutation-based analysis.

Permutation-based feature importance was calculated from the Random Forest model fitted to the full dataset, and the resulting rankings are presented in Table 3. The six highest-ranked features from the exploratory Random Forest permutation-importance analysis were used to con-

struct a simplified decision tree for descriptive visualisation only. Because both feature selection and tree construction were performed using the same dataset, the resulting structure and thresholds are data-driven and potentially optimistic. No independent, cross-validated, or externally validated discrimination or calibration was estimated for this tree. The decision tree is therefore presented only in the **Supplementary Material** and should not be interpreted as a validated predictive model or as providing clinically actionable cut-offs. SHapley Additive exPlanations (SHAP) analysis was not performed in the present study.

The dataset consisted of 133 patients, of whom 13 (9.8%) exhibited ISUP Grade Group ≥3 upgrading, resulting in substantial class imbalance. No synthetic oversampling, undersampling, Synthetic Minority Over-sampling Technique (SMOTE), or class-weighting strategy was applied in the primary analysis. Given the small number of outcome events, we elected to preserve the observed event distribution and avoid introducing an additional resampling or weighting step that could further increase modelling variability. Instead, stratification was used during repeated cross-validation to preserve the outcome distribution across folds. Model performance was evaluated using the area under the receiver operating characteristic curve (ROC-AUC) together with the area under the precision–recall curve (PR-AUC) and the Brier score, with PR-AUC included to provide additional information on performance for the minority outcome class. This approach does not eliminate the potential effects of class imbalance, and all model-performance estimates were therefore considered exploratory.

Internal validation procedures, including repeated stratified cross-validation and bootstrap resampling, are described in the "Dataset Splitting and Validation" subsection.

A total of 11 supervised ML algorithms were evaluated using the same cross-validation partitions. The evaluated classifiers were Logistic Regression, Random Forest, linear support vector machine (SVM), radial basis function support vector machine (RBF SVM), K-Nearest Neighbours (KNN), Gaussian Naive Bayes, Decision Tree, Gaussian Process Classifier, AdaBoost, Artificial Neural Networks (multi-layer perceptron [MLP]), and Quadratic Discriminant Analysis (QDA). Logistic Regression was considered the conventional linear baseline model, whereas the remaining algorithms were evaluated to determine whether nonlinear or ensemble-based approaches could achieve better discrimination.

Model performance was evaluated using the ROC-AUC, which quantifies the model's ability to discriminate between patients with and without ISUP Grade Group ≥3 upgrading across all possible classification thresholds. The receiver operating characteristic (ROC) curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at varying decision thresholds. These metrics were defined as follows:

$$\text{TPR (sensitivity)} = TP / (TP + FN)$$

$$\text{FPR (1 – specificity)} = FP / (FP + TN)$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. ROC-AUC represents the probability that the model assigns a higher predicted score to a randomly chosen positive case than to a randomly chosen negative case. An AUC of 0.5 indicates no discriminative ability (random classification), whereas a value of 1.0 indicates perfect discrimination.

ROC-AUC was selected as the primary discrimination metric because it summarises the model's ability to distinguish patients with and without ISUP Grade Group ≥3 upgrading across all possible classification thresholds, thereby reducing dependence on a single preselected cut-off. However, threshold independence does not compensate for the limited number of outcome events or guarantee the stability of performance estimates. PR-AUC was calculated from the positive-class predicted probabilities by constructing the precision–recall curve using the precision_recall_curve function in scikit-learn. The area under the resulting precision–recall curve was calculated using trapezoidal integration. PR-AUC was evaluated on the corresponding validation data and summarised as the mean ± standard deviation across the repeated cross-validation procedure. Given the limited event count, all performance estimates and model comparisons were considered exploratory.

For Brier-score calculation, predicted probabilities for the positive outcome (postoperative ISUP Grade Group ≥3) were obtained using the fitted classifier's (predict_proba) method. The Brier score was calculated by comparing these probability estimates with the corresponding binary outcome indicators (1 for ISUP Grade Group ≥3 and 0 otherwise). Thus, Brier scores were calculated from predicted probabilities rather than binary class predictions or decision-function scores. The preprocessing and classifier were implemented as a single pipeline, with preprocessing fitted on the training data and subsequently applied to the corresponding validation data before probability estimation.

Hyperparameter optimisation was performed for Random Forest, Gaussian Naïve Bayes, linear SVM, RBF SVM, K-Nearest Neighbours, Decision Tree, AdaBoost, multi-layer perceptron, and Quadratic Discriminant Analysis. Logistic Regression and Gaussian Process Classifier were evaluated using prespecified fixed configurations. For the tuned classifiers, hyperparameter selection was performed separately within each outer training fold using exhaustive grid search and stratified 3-fold inner cross-validation with shuffling. ROC-AUC was used as the optimisation criterion, and the hyperparameter configuration yielding the highest mean inner-validation ROC-AUC was selected. The selected configuration was subsequently refitted on the complete outer training fold and evaluated on the corresponding outer validation fold. All preprocessing steps were restricted to the training data to prevent infor-

mation leakage. Model-specific fixed settings and hyperparameter search spaces are provided in **Supplementary Table 1**.

ML models were developed using Python programming language version 3.10 (Python Software Foundation; https://www.python.org/) and PyCharm IDE version 2024.3 (JetBrains; https://www.jetbrains.com/pycharm/whatsnew/2024-3/). The ML algorithms were implemented using the Scikit-learn library version 1.6 (scikit-learn developers) [19]. The results of the applied models were comparatively analysed for predicting ISUP Grade Group ≥3 upgrading in prostate cancer diagnosis. Through this approach, the impact of demographic, clinical, and pathological factors on predicting ISUP Grade Group ≥3 outcomes was thoroughly examined.

Dataset Splitting and Validation

Because only 13 of 133 patients (9.8%) exhibited ISUP Grade Group ≥3 upgrading, model performance was assessed using repeated stratified cross-validation in order to reduce the instability associated with a single random train–test split. The dataset was evaluated using repeated stratified 5-fold cross-validation with 10 repetitions, thereby preserving the class distribution in each fold and providing an assessment of performance variability across repeated data partitions.

Hyperparameter optimisation, when applicable, was performed within the outer training folds using the nested cross-validation procedure described in the "ML-Based Classification Methods" subsection and **Supplementary Table 1**. The repeated stratified cross-validation results were used as the primary internal-validation estimates and are presented in Table 2.

As a complementary internal-validation analysis, bootstrap resampling was performed for the Random Forest model using 1000 bootstrap repetitions. The individual patient was used as the sampling unit. In each iteration, 133 patients were sampled with replacement from the complete-case cohort. The complete modelling pipeline, including preprocessing and hyperparameter selection, was repeated within each bootstrap sample. ROC-AUC was calculated from continuous positive-class prediction scores. PR-AUC was calculated from positive-class predicted probabilities by constructing the precision–recall curve using the precision_recall_curve function and calculating the area under the resulting curve using trapezoidal integration. Bootstrap distributions were summarised using the median, and 95% uncertainty intervals were calculated using the percentile method, corresponding to the 2.5th and 97.5th percentiles of the bootstrap distribution. Bootstrap samples containing only one outcome class were discarded and redrawn. Optimism correction was performed for ROC-AUC by calculating, in each bootstrap iteration, the difference between performance in the bootstrap sample and performance obtained by applying the bootstrap-fitted model to the original cohort. Mean optimism across the 1000 iterations was subtracted from the apparent ROC-AUC of the Random Forest model fitted to the complete cohort to obtain the optimism-corrected ROC-AUC. The bootstrap-based robustness results are presented in **Supplementary Table 2**.

All preprocessing steps, including standardisation, were fitted within each training fold and subsequently applied only to the corresponding validation fold. Because ROC-AUC estimates obtained across repeated cross-validation folds are not statistically independent, variability in model performance is summarised descriptively using the mean and standard deviation across repeated validation folds. Formal confidence intervals based on treating fold-level estimates as independent observations are not reported. This procedure was used to minimise data leakage and to provide less biased estimates of model discrimination and overall performance.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA). The distribution of continuous variables was re-assessed using the Shapiro–Wilk test. Normally distributed variables are presented as the mean ± standard deviation and were compared using the independent-samples Student's *t*-test, whereas non-normally distributed variables are presented as the median and interquartile range and were compared using the Mann–Whitney U test. For reporting in Table 1, test statistics are presented as t values for the Student's *t*-test and standardised Z values for the Mann–Whitney U test. Biopsy core counts were treated as quantitative discrete variables, whereas positive biopsy core percentage and maximum and minimum tumour percentages per core were treated as quantitative continuous variables. These variables were analysed using their original numerical values because no prespecified clinically validated cut-points were available. Categorical variables are expressed as numbers and percentages and were compared using Fisher's exact test or the Fisher–Freeman–Halton exact test with Monte Carlo estimation, as appropriate. No separate test statistic is reported for exact tests. A p-value < 0.05 was considered statistically significant. These statistical analyses were used to support the interpretation of the ML findings.

Results

Patient Characteristics and Conventional Statistical Comparisons

A total of 133 patients were included in the final analysis, of whom 13 (9.8%) exhibited major upgrading to ISUP Grade Group ≥3. Baseline demographic, clinical, and pathological characteristics, together with the corresponding test statistics and p-values, are summarised in Table 1. Patients with major upgrading had a higher median BMI

(27.4 vs 25.5 kg/m^2; p = 0.005), a lower median prostate volume (35.0 vs 64.5 mL; p = 0.030), and a higher minimum tumour percentage per core (10.0% vs 5.0%; p = 0.024). Perineural invasion and bilateral tumour involvement were numerically more frequent in the upgrading group; however, these differences did not reach statistical significance (p = 0.076 and p = 0.061, respectively). The clinical T-stage distribution differed significantly between the groups (p = 0.043).

Comparative analysis revealed statistically significant differences between patients with and without major upgrading in BMI (p = 0.005), prostate volume (p = 0.030), clinical T-stage distribution (p = 0.043), and minimum tumour percentage per core (p = 0.024). However, there were no significant differences in age (p = 0.961), perineural invasion (p = 0.076), bilateral tumour involvement (p = 0.061), hypertension, diabetes mellitus, PSA, PSA density, number of biopsy cores, percentage or number of positive biopsy cores, maximum tumour percentage per core, or number of tumour-infiltrated biopsy areas (all remaining $p > 0.05$).

ML Performance and Internal Validation

Table 2 summarises the performance of all 11 evaluated ML models under repeated stratified 5-fold cross-validation with 10 repetitions, including the mean ROC-AUC and mean PR-AUC with the standard deviations as well as the Brier score. Random Forest yielded the highest mean ROC-AUC (0.83 ± 0.09), followed by Gaussian Naïve Bayes (0.78 ± 0.10). Compared with the Logistic Regression baseline (mean ROC-AUC, 0.74 ± 0.11), the absolute numerical difference in mean ROC-AUC was 0.09; however, this difference was not interpreted as evidence of model superiority. Given the limited number of outcome events, these differences should be interpreted descriptively, considered exploratory, and do not establish definitive model superiority.

As a complementary robustness assessment, bootstrap resampling was performed for the Random Forest model, which had the highest mean ROC-AUC under repeated cross-validation. The bootstrap-based results are presented in **Supplementary Table 2**. Given the limited number of upgrading events, these bootstrap estimates should be interpreted as exploratory and complementary to the primary repeated cross-validation results.

Exploratory Feature Importance and Decision Tree Analyses

Permutation-based feature importance analysis ranked prostate volume, the number of biopsy cores, and the percentage of positive biopsy cores as the three highest-ranked features. The remaining variables showed lower feature-importance scores within the present dataset (Table 3). Because the model and feature importance analysis used the full dataset, these rankings should be regarded as descriptive and exploratory and should not be interpreted as independently validated predictor effects, statistical significance, causal associations, or clinically validated importance. The corresponding permutation-based feature-importance scores are visually presented in Fig. 2, illustrating the relative contribution of the evaluated variables within the Random Forest model.

A simplified decision tree based on the six highest-ranked features was generated solely as an exploratory descriptive visualisation and is provided in **Supplementary Fig. 1**. Because the same dataset was used for feature ranking and tree construction, the resulting structure and thresholds should be considered data-driven and potentially unstable.

Discussion

The present study explored the ability of multiple ML approaches to predict ISUP Grade Group ≥3 upgrading after radical prostatectomy in patients initially classified as having D'Amico low-risk prostate cancer. Under repeated stratified 5-fold cross-validation, Random Forest yielded the highest mean ROC-AUC (0.83 ± 0.09), whereas Logistic Regression, used as the conventional linear baseline classifier, yielded a mean ROC-AUC of 0.74 ± 0.11. Although the absolute numerical difference in mean ROC-AUC was 0.09, the limited number of upgrading events and substantial variability across repeated validation folds preclude a definitive conclusion regarding model superiority. Therefore, the present findings should be interpreted as an exploratory comparison of predictive approaches rather than evidence of the clinical superiority of Random Forest.

The conventional group comparisons identified several demographic, clinical, and biopsy-related characteristics that differed between patients with and without major upgrading. Age was similar between the two groups in the present cohort (p = 0.961), although Wang et al. [20] reported an increased risk of Gleason score upgrading and pathological upstaging with increasing age. PSA density was numerically higher in patients with major upgrading (0.123 vs 0.096), but this difference did not reach statistical significance (p = 0.096), whereas Maruyama et al. [21] reported PSA density as a factor associated with pathological upgrading. In contrast, BMI was significantly higher in the upgrading group in our cohort (27.4 vs 25.5 kg/m^2; p = 0.005), which is directionally consistent with the findings reported by Takeshima et al. [22] regarding BMI and ISUP grade upgrading. Differences between the studies may reflect variation in patient selection, biopsy practices, outcome definitions, and cohort composition. Importantly, these conventional comparisons in our study were univariable and should not be interpreted as independently validated predictor effects.

ML has increasingly been investigated as a method for integrating multiple clinical and pathological variables in prostate cancer risk stratification. Tward et al. [23]

Table 1. Demographic, clinical, and pathological characteristics.

Variable	Major upgrading to ISUP Grade Group ≥3, Yes (n = 13)	No major upgrading, ISUP Grades 1–2 (n = 120)	Test statistic	*p*-value
Age (years)	62.4 ± 8.0	62.3 ± 6.8	$t = 0.050$	0.961[1]
Body mass index (BMI) (kg/m^2)	27.4 (25.2–29.3)	25.5 (23.7–29.1)	Z = 2.807	0.005[2]
Hypertension, n (%)				
No	12 (92.3)	81 (67.5)		0.108[3]
Yes	1 (7.7)	39 (32.5)	–	
Diabetes mellitus, n (%)				
No	13 (100.0)	104 (86.7)		0.365[3]
Yes	0 (0.0)	16 (13.3)		
Prostate volume (mL)	35.0 (28.0–67.0)	64.5 (41.0–87.0)	Z = 2.170	0.030[2]
PSA (ng/mL)	5.4 (4.3–6.7)	6.2 (4.9–8.0)	Z = 0.902	0.367[2]
PSA density	0.123 (0.095–0.239)	0.096 (0.061–0.146)	Z = 1.665	0.096[2]
Perineural invasion, n (%)				
No	10 (76.9)	112 (93.3)		0.076[3]
Yes	3 (23.1)	8 (6.7)	–	
Clinical T stage, n (%)				
T1a	0 (0.0)	2 (1.7)		
T1b	2 (15.4)	6 (5.0)	–	0.043[4]
T1c	8 (61.5)	42 (35.0)		
T2a	3 (23.1)	70 (58.3)		
Number of biopsy cores	8.0 (6.0–12.0)	12.0 (8.0–16.2)	Z = 1.799	0.072[2]
Percentage of positive biopsy cores (%)	37.5 (27.3–50.0)	30.0 (16.7–41.7)	Z = 1.522	0.128[2]
Number of positive biopsy cores	4.0 (3.0–5.0)	3.0 (2.0–5.0)	Z = 0.513	0.608[2]
Maximum tumour percentage per core (%)	40.0 (30.0–70.0)	30.0 (10.0–50.0)	Z = 1.296	0.195[2]
Minimum tumour percentage per core (%)	10.0 (5.0–20.0)	5.0 (0.0–10.0)	Z = 2.257	0.024[2]
Bilateral tumour involvement, n (%)				
No	5 (38.5)	82 (68.3)	–	0.061[3]
Yes	8 (61.5)	38 (31.7)		
Number of tumour-infiltrated areas	3.0 (2.0–4.0)	2.0 (1.0–3.0)	Z = 1.717	0.086[2]

[1] The data are presented as mean ± standard deviation and were compared using the independent-samples Student's *t*-test.

[2] The data are presented as median (interquartile range) and were compared using the Mann–Whitney U test.

[3] The data are presented as n (%) and were compared using Fisher's exact test.

[4] Clinical T stage was analysed as a categorical variable for the between-group comparison using the Fisher–Freeman–Halton exact test with Monte Carlo estimation. The *p*-value refers to the overall comparison of the T-stage distribution, not to individual categories or a linear trend across categories. This categorical comparison was separate from the integer coding used in the machine learning models, as described in the "ML-Based Classification Methods" subsection.

– indicates that no separate test statistic is reported for exact tests.

Abbreviations: ISUP, International Society of Urological Pathology; PSA, prostate-specific antigen.

demonstrated the potential of multimodal deep learning approaches for prostate cancer risk stratification using digital histopathology, illustrating the broader role of ML methods in this setting. Nevertheless, the clinical value of such models depends strongly on the population studied, the available variables, the number of outcome events, and the validation strategy. In the present study, routinely available preoperative clinical and biopsy-derived variables were deliberately evaluated, making the approach potentially reproducible in conventional clinical datasets, although external validation remains necessary.

Previous ML studies specifically addressing the Gleason score or ISUP grade upgrading have reported heterogeneous predictors and different best-performing algorithms. Liu et al. [24] reported that Lasso logistic regression achieved the best performance in their cohort and identified PSA, biopsy Gleason Grade Group, apical involvement, and clinical T stage among the relevant predictors. ElKarami et al. [25] used magnetic resonance imaging (MRI)-targeted biopsy data in patients eligible for active surveillance and identified age, BMI, and positive biopsy cores among the important variables associated with upgrading.

Table 2. Performance of all evaluated machine learning models under repeated stratified 5-fold cross-validation with 10 repetitions.

Model	ROC-AUC, mean ± SD	PR-AUC, mean ± SD	Brier score
Random Forest	0.83 ± 0.09	0.34 ± 0.10	0.085
Gaussian Naïve Bayes	0.78 ± 0.10	0.28 ± 0.09	0.094
Linear support vector machine	0.75 ± 0.11	0.24 ± 0.08	0.101
Logistic Regression	0.74 ± 0.11	0.23 ± 0.07	0.103
Neural Network	0.73 ± 0.13	0.22 ± 0.09	0.108
AdaBoost	0.72 ± 0.12	0.21 ± 0.08	0.110
Gaussian Process Classifier	0.63 ± 0.21	0.30 ± 0.19	0.216
Decision Tree	0.53 ± 0.13	0.16 ± 0.15	0.182
Quadratic Discriminant Analysis	0.51 ± 0.22	0.21 ± 0.14	0.123
Radial basis function support vector machine	0.45 ± 0.15	0.16 ± 0.08	0.091
K-Nearest Neighbours	0.42 ± 0.11	0.10 ± 0.02	0.109

Abbreviations: PR-AUC, area under the precision–recall curve; ROC-AUC, area under the receiver operating characteristic curve; SD, standard deviation.

Note: Given the low outcome prevalence (13/133, 9.8%), PR-AUC is reported alongside ROC-AUC to provide complementary information on performance for the minority outcome class. No class rebalancing was applied in the primary analysis.

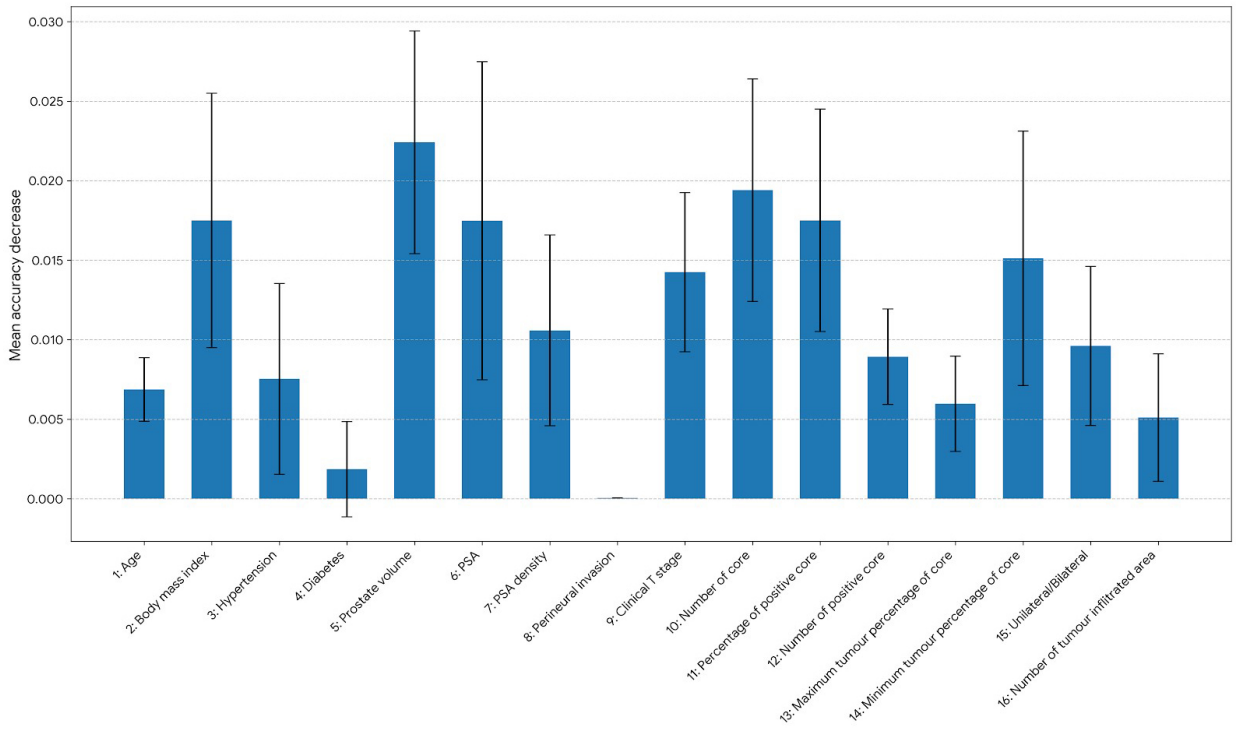


Fig. 2. Exploratory permutation-based feature importance analysis using random forest. Feature importance scores were calculated from the Random Forest model fitted to the full dataset. These scores describe the relative contribution of the variables within the present dataset. Because the same dataset was used for model fitting and feature-importance estimation, the findings should be considered exploratory and should not be interpreted as independently validated effects or statistical significance.

Li et al. [26] compared nine machine learning models for predicting upgrading from biopsy to radical prostatectomy pathology. LightGBM achieved the highest test-set ROC-AUC (0.845), with an ROC-AUC of 0.766 in external validation. The selected model inputs included biopsy ISUP Grade Group, age, clinical T stage, BMI, and PSA. Taken together, these studies indicate that no single algorithm or predictor profile has been consistently dominant across cohorts. Differences in population characteristics, feature availability, outcome prevalence, and validation methodology probably contribute to this heterogeneity.

Table 3. Exploratory permutation-based feature-importance rankings from the random forest model fitted to the full dataset.

#	Feature	Score
1	Prostate volume	0.022418
2	Number of biopsy cores	0.019405
3	Percentage of positive biopsy cores	0.017506
4	Body mass index	0.017494
5	Prostate-specific antigen	0.017479
6	Minimum tumour percentage per core	0.015113
7	Clinical T stage†	0.014247
8	Prostate-specific antigen density	0.010582
9	Bilateral tumour involvement	0.009611
10	Number of positive biopsy cores	0.008927
11	Hypertension	0.007540
12	Age	0.006866
13	Maximum tumour percentage per core	0.005970
14	Number of tumour-infiltrated biopsy areas	0.005102
15	Diabetes mellitus	0.001856
16	Perineural invasion	0.000045

Note: †Clinical T stage was entered into the Random Forest model as a single integer-coded feature (T1a = 1, T1b = 2, T1c = 3, and T2a = 4), rather than using dummy/one-hot encoding. Its permutation-importance score describes its contribution within the fitted model under this encoding and should be interpreted as exploratory. This score does not establish an independent predictor effect or indicate the direction of the association with upgrading. Robustness to alternative categorical encoding has not been established.

In the present exploratory Random Forest feature importance analysis, prostate volume, the number of biopsy cores, the percentage of positive biopsy cores, BMI, PSA, and the minimum tumour percentage per core were among the highest-ranked variables. The relatively high ranking of the number of biopsy cores should be interpreted cautiously. Biopsy technique was not included as an input feature in the original modelling framework, although both cognitive TRUS-guided and MR-fusion biopsies were represented in the cohort. Because the number of biopsy cores varied according to biopsy approach and clinical practice over the long study period, this variable may partly capture differences related to biopsy technique, diagnostic strategy, or treatment era rather than an independent biological relationship with upgrading. In addition, changes in biopsy practice over time may have introduced residual confounding that could not be separated from the effect of biopsy-core number in the present dataset. Given the limited number of upgrading events, stratified or sensitivity analyses according to biopsy technique were not considered sufficiently reliable. Therefore, the feature importance of biopsy-core number should be regarded as exploratory and requires confirmation in larger cohorts with standardised contemporary biopsy protocols. Clinical T stage ranked seventh in the exploratory permutation-based feature-importance analysis. This ranking describes its contribution within the fitted Random Forest model under the integer coding used and should not be interpreted as evidence of an independent predictor effect or established clinical importance. Conventional categorical comparisons also showed a difference in T-stage distribution between the outcome groups; however, this does not establish a monotonic risk gradient across T1a, T1b, T1c, and T2a or validate their equally spaced numerical coding. Björnebo et al. [27] previously identified clinical T stage among factors associated with adverse pathological findings in men initially managed with active surveillance for low-risk prostate cancer. Nevertheless, the present T-stage findings remain exploratory, and their robustness to alternative categorical encoding has not been assessed.

An important finding was that the variables identified by conventional statistical comparisons did not completely overlap with those ranked highly by the Random Forest model. For example, perineural invasion and bilateral tumour involvement were numerically more frequent in patients with major upgrading but did not reach statistical significance in the conventional group comparisons ($p = 0.076$ and $p = 0.061$, respectively) and ranked relatively low in the permutation-based feature importance analysis. Conversely, some biopsy core–related variables ranked relatively highly in the ML analysis despite not demonstrating statistically significant differences in the univariable group comparisons. These findings are not necessarily contradictory because the two analytical approaches address different questions. Conventional comparisons assess individual variables between outcome groups, whereas permutation importance evaluates the contribution of each variable to model performance in the context of the other included features. However, because the Random Forest model and permutation-based feature importance analysis were performed using the same full dataset, the resulting rankings should be regarded as descriptive and exploratory rather than as evidence of statistical significance, causality, or independently validated predictor effects.

Several strengths of the present study should be considered together with its limitations. Multiple supervised ML algorithms were evaluated within the same cohort, and repeated stratified 5-fold cross-validation with 10 repetitions was used as the primary internal-validation approach. Bootstrap resampling was additionally used as a complementary robustness assessment. Together, these approaches allowed model performance to be assessed across repeated resampling procedures rather than relying on a single arbitrary train–test partition. However, they do not increase the number of independent outcome events. Only 13 of 133 patients (9.8%) experienced ISUP Grade Group ≥3 upgrading, which substantially limits the effective information available for model development. Consequently, ROC-AUC estimates, model rankings, feature-importance

scores, and other model-derived findings may remain unstable despite repeated internal validation. Class imbalance represents an additional modelling limitation. Because only 13 patients experienced the outcome, classifiers may preferentially fit the majority class, and performance for the minority class may be unstable even when stratified cross-validation is used. In particular, apparently acceptable ROC-AUC values may coexist with limited precision–recall performance in an imbalanced dataset. No class-weighted or resampling-based sensitivity analysis was performed; therefore, the extent to which alternative imbalance-aware strategies would affect model rankings and performance estimates remains unknown. Future studies with larger numbers of outcome events should evaluate class-weighted and other imbalance-aware approaches within the same internal-validation framework.

The retrospective, single-centre design represents an additional limitation of this study and restricts the generalisability of the findings. In addition, the study cohort consisted exclusively of patients who underwent radical prostatectomy. Patients with low-risk prostate cancer managed with active surveillance or other non-surgical approaches were not represented, and selection into the surgical cohort may therefore have introduced additional selection bias. Consequently, the present findings should not be directly extrapolated to the entire population of patients with low-risk prostate cancer. Twenty-six otherwise eligible patients were excluded because of missing data in at least one candidate variable, and the use of complete-case analysis may have introduced selection bias if missingness was not random. Multiparametric MRI data were not consistently available throughout the study period and therefore could not be systematically incorporated into the predictive models. Biopsy technique and clinical biopsy practice also changed over the study period, and biopsy technique was not included in the original candidate predictor set. Residual confounding by biopsy approach and treatment era therefore cannot be excluded. Given the limited number of upgrading events, this study was not sufficiently powered for reliable stratified or sensitivity analyses according to biopsy technique. Furthermore, pathology specimens were not re-reviewed specifically for this study. The analyses were based on finalised routine pathology reports evaluated by a single uropathologist, and because no second independent pathological review was performed, interobserver agreement could not be assessed. The long study period also spanned changes in prostate cancer grading recommendations and reporting practices, including the introduction and subsequent refinement of the ISUP Grade Group system [5,28]. Although historical Gleason scores were retrospectively mapped to the corresponding Grade Groups to provide a common reporting framework, this harmonisation does not constitute histopathological re-review and cannot account for possible temporal changes in the interpretation of Gleason patterns [5]. Consequently, temporal variation in pathological grading practice may have influenced baseline classification, the recorded frequency of major upgrading, and the resulting model estimates. The magnitude and direction of this potential effect cannot be determined from the present data.

Several modelling limitations should also be emphasised. Clinical T stage was represented by ordered integer values rather than dummy/one-hot encoding. In logistic regression, this specification imposes a linear contribution of T stage to the log-odds of upgrading, whereas in SVM and KNN it may influence the numerical geometry or distances used by the models. Because alternative encoding was not evaluated, the extent to which this modelling choice influenced predictive performance or the feature importance of T stage remains unknown. Future analyses should assess the robustness of these findings using categorical encoding. Reliable calibration could not be established in the present low-event dataset. Calibration intercept and slope estimates could not be reliably interpreted because of the limited number of outcome events, the absence of confidence intervals, and the inability to fully verify the source of the predictions used for their calculation. In addition, the prediction source underlying the original calibration curve could not be fully verified from the retained analysis outputs. Consequently, the calibration intercept and slope estimates and the corresponding calibration curve were removed and were not retained as evidence of model reliability. Nested cross-validation was used for models requiring hyperparameter optimisation but was not applied uniformly to all classifiers. In addition, neither penalised/Firth logistic regression nor an a priori reduction of the candidate predictor set was performed. Finally, the use of the same dataset for Random Forest feature ranking and subsequent construction of the simplified decision tree introduces repeated data use and may result in optimistic and unstable tree structures and thresholds. For this reason, the tree was retained only as an exploratory visualisation in the **Supplementary Material**, and its split points should not be interpreted as validated clinical cut-offs.

Taken together, the present findings suggest that routinely available clinical and biopsy-derived variables may contain useful information for preoperative risk assessment among selected patients with D'Amico low-risk prostate cancer undergoing or being considered for radical prostatectomy. However, the current results do not establish independent predictor effects, clinically validated thresholds, or definitive superiority of one ML algorithm over another. External validation in substantially larger, preferably multicentre cohorts with a greater number of upgrading events, standardised pathological assessment, contemporary imaging information, and robust evaluation of both discrimination and calibration is required before clinical application.

Conclusion

This exploratory study suggests that ML methods may contribute to the prediction of ISUP Grade Group ≥3 upgrading among surgically treated patients with D'Amico low-risk prostate cancer. Random Forest yielded the highest mean ROC-AUC under repeated stratified cross-validation. These findings support further evaluation of ML-based risk stratification in larger, independent multicentre cohorts before clinical application.

Abbreviations

AI, artificial intelligence; AUA, American Urological Association; BMI, body mass index; EAU, European Association of Urology; FN, false negative; FP, false positive; FPR, false positive rate; ISUP, International Society of Urological Pathology; KNN, K-Nearest Neighbours; ML, machine learning; MLP, multi-layer perceptron; MR, magnetic resonance; MRI, magnetic resonance imaging; NCCN, National Comprehensive Cancer Network; PR-AUC, area under the precision–recall curve; PSA, prostate-specific antigen; QDA, Quadratic Discriminant Analysis; RBF SVM, radial basis function support vector machine; ROC, receiver operating characteristic; ROC-AUC, area under the receiver operating characteristic curve; SHAP, SHapley Additive exPlanations; SMOTE, Synthetic Minority Over-sampling Technique; SVM, support vector machine; TN, true negative; TP, true positive; TPR, true positive rate; TRUS, transrectal ultrasound.

Availability of Data and Materials

The datasets generated or analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions

IOY: Conceived the study, supervised the research process, and provided methodological guidance. OY: Contributed to the development and implementation of machine learning models, participated in data interpretation, and reviewed the results for scientific accuracy. MZ: Performed statistical analyses, managed data documentation, and supported statistical data interpretation. RK: Led the selection and optimization of machine learning algorithms, participated in model development, and contributed to the interpretation of results. NA: Participated in the study design and coordinated data collection. MGA: Conducted the literature review and data curation. MD: Contributed to statistical analyses and data evaluation. VI: Made substantial contributions to the acquisition of clinical data and to the analysis and interpretation of the study results, and provided senior supervision. MZ and MGA have been involved in drafting the manuscript and all authors have been involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

Ethical approval was obtained from the Çukurova University Non-Interventional Clinical Research Ethics Committee (Approval No: 149/68; 8 November 2024). The study complied with the Declaration of Helsinki. The requirement for informed consent was waived by the Çukurova University Non-Interventional Clinical Research Ethics Committee due to the retrospective design, and there was no contact with patients.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.24976/Discov.Med.202638214.278.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021; 71: 209–249. https://doi.org/10.3322/caac.21660

[2] ECIS - European Cancer Information System. Prostate cancer burden in EU-27. 2021. Available at: https://ecis.jrc.ec.europa.eu/sites/default/files/2023-12/prostate_cancer_En-Nov_2021.pdf (Accessed: 1 May 2026).

[3] Cornford P, van den Bergh RCN, Briers E, Van den Broeck T, Brunckhorst O, Darraugh J, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer. EAU Guidelines Office: Arnhem, The Netherlands. 2024.

[4] D'Amico AV, Whittington R, Malkowicz SB, Cote K, Loffredo M, Schultz D, et al. Biochemical outcome after radical prostatectomy or external beam radiation therapy for patients with clinically localized prostate carcinoma in the prostate specific antigen era. Cancer. 2002; 95: 281–286. https://doi.org/10.1002/cncr.10657

[5] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urologi-

cal Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. The American Journal of Surgical Pathology. 2016; 40: 244–252. https://doi.org/10.1097/PAS.0000000000000530

[6] Novosel M, Smani S, Marks VA, Jeong F, Sprenkle P, Leapman M. Associations between patient sociodemographic factors and non-treatment for localized prostate cancer. Journal of Clinical Oncology. 2023; 41: 307–307. https://doi.org/10.1200/jco.2023.41.6_suppl.307

[7] Schaeffer EM, Srinivas S, Adra N, An Y, Barocas D, Bitting R, et al. Prostate Cancer, Version 4.2023, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network: JNCCN. 2023; 21: 1067–1096. https://doi.org/10.6004/jnccn.2023.0050

[8] Smith JA, Jr. Radical prostatectomy for low risk carcinoma of the prostate. World Journal of Urology. 2008; 26: 443–446. https://doi.org/10.1007/s00345-008-0293-9

[9] Zou Q, Cao J, Chen Z, Wang S, Gu C, Li S, et al. Prostate Volume is A Predictor of Gleason Score Upgrading after Radical Prostatectomy in Low-Risk Prostate Cancer: A Systematic Review and Meta-analysis. Urology Journal. 2024; 21: 20–28. https://doi.org/10.22037/uj.v20i.7796

[10] Porcaro AB, Bianchi A, Panunzio A, Gallina S, Tafuri A, Serafin E, et al. The impact of prognostic group classification on prostate cancer progression in intermediate-risk patients according to the European Association of Urology system: results in 479 patients treated with robot-assisted radical prostatectomy at a single tertiary referral center. Therapeutic Advances in Urology. 2024; 16: 17562872241229260. https://doi.org/10.1177/17562872241229260

[11] Nguyen DD, Haeuser L, Paciotti M, Reitblat C, Cellini J, Lipsitz SR, et al. Systematic Review of Time to Definitive Treatment for Intermediate Risk and High Risk Prostate Cancer: Are Delays Associated with Worse Outcomes? Journal of Urology. 2021; 205: 1263–1274. https://doi.org/10.1097/ju.0000000000001601

[12] Porcaro AB, Siracusano S, de Luyk N, Corsi P, Sebben M, Tafuri A, et al. Low-Risk Prostate Cancer and Tumor Upgrading in the Surgical Specimen: Analysis of Clinical Factors Predicting Tumor Upgrading in a Contemporary Series of Patients Who were Evaluated According to the Modified Gleason Score Grading System. Current Urology. 2017; 10: 118–125. https://doi.org/10.1159/000447164

[13] Schiffmann J, Wenzel P, Salomon G, Budäus L, Schlomm T, Minner S, et al. Heterogeneity in D'Amico classification-based low-risk prostate cancer: Differences in upgrading and upstaging according to active surveillance eligibility. Urologic Oncology. 2015; 33: 329.e13–329.e3.29E19. https://doi.org/10.1016/j.urolonc.2015.04.004

[14] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017; 42: 60–88. https://doi.org/10.1016/j.media.2017.07.005

[15] Morozov A, Taratkin M, Bazarkin A, Rivas JG, Puliatti S, Checcucci E, et al. A systematic review and meta-analysis of artificial intelligence diagnostic accuracy in prostate cancer histology identification and grading. Prostate Cancer and Prostatic Diseases. 2023; 26: 681–692. https://doi.org/10.1038/s41391-023-00673-3

[16] Wang H, Xia Z, Xu Y, Sun J, Wu J. The predictive value of machine learning and nomograms for lymph node metastasis of prostate cancer: a systematic review and meta-analysis. Prostate Cancer and Prostatic Diseases. 2023; 26: 602–613. https://doi.org/10.1038/s41391-023-00704-z

[17] Castaldo R, Cavaliere C, Soricelli A, Salvatore M, Pecchia L, Franzese M. Radiomic and Genomic Machine Learning Method Performance for Prostate Cancer Diagnosis: Systematic Literature Review. Journal of Medical Internet Research. 2021; 23: e22394. https://doi.org/10.2196/22394

[18] Ozbozduman K, Loc I, Durmaz S, Atasoy D, Kilic M, Yildirim H, et al. Machine learning prediction of Gleason grade group upgrade between in-bore biopsy and radical prostatectomy pathology. Scientific Reports. 2024; 14: 5849. https://doi.org/10.1038/s41598-024-56415-5

[19] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[20] Wang X, Zhang Y, Ji Z, Yang P, Tian Y. Old men with prostate cancer have higher risk of Gleason score upgrading and pathological upstaging after initial diagnosis: a systematic review and meta-analysis. World Journal of Surgical Oncology. 2021; 19: 18. https://doi.org/10.1186/s12957-021-02127-3

[21] Maruyama Y, Sadahira T, Araki M, Mitsui Y, Wada K, Rodrigo A, et al. Factors predicting pathological upgrading after prostatectomy in patients with Gleason grade group 1 prostate cancer based on opinion-matched biopsy specimens. Molecular and Clinical Oncology. 2020. https://doi.org/10.3892/mco.2020.1996

[22] Takeshima Y, Yamada Y, Teshima T, Fujimura T, Kakutani S, Hakozaki Y, et al. Clinical significance and risk factors of International Society of Urological Pathology (ISUP) grade upgrading in prostate cancer patients undergoing robot-assisted radical prostatectomy. BMC Cancer. 2021; 21: 501. https://doi.org/10.1186/s12885-021-08248-y

[23] Tward JD, Huang HC, Esteva A, Mohamad O, van der Wal D, Simko JP, et al. Prostate Cancer Risk Stratification in NRG Oncology Phase III Randomized Trials Using Multimodal Deep Learning With Digital Histopathology. JCO Precision Oncology. 2024; 8: e2400145. https://doi.org/10.1200/PO.24.00145

[24] Liu H, Tang K, Peng E, Wang L, Xia D, Chen Z. Predicting Prostate Cancer Upgrading of Biopsy Gleason Grade Group at Radical Prostatectomy Using Machine Learning-Assisted Decision-Support Models. Cancer Management and Research. 2020; 12: 13099–13110. https://doi.org/10.2147/CMAR.S286167

[25] ElKarami B, Deebajah M, Polk S, Peabody J, Shahrrava B, Menon M, et al. Machine learning-based prediction of upgrading on magnetic resonance imaging targeted biopsy in patients eligible for active surveillance. Urologic Oncology. 2022; 40: 191.e15–191.e20. https://doi.org/10.1016/j.urolonc.2022.01.012

[26] Li SF, Zhao JG, Jiang CY, Wang SY, Liu SY, Zhang YJ, et al. Development and validation of an interpretable machine learning model for predicting Gleason score upgrade in prostate cancer. Translational Andrology and Urology. 2025; 14: 1631–1644. https://doi.org/10.21037/tau-2025-178

[27] Björnebo L, Olsson H, Nordström T, Jäderling F, Grönberg H, Eklund M, et al. Predictors of adverse pathology on radical prostatectomy specimen in men initially enrolled in active surveillance for low-risk prostate cancer. World Journal of Urology. 2021; 39: 1797–1804. https://doi.org/10.1007/s00345-020-03394-7

[28] van Leenders GJLH, van der Kwast TH, Grignon DJ, Evans AJ, Kristiansen G, Kweldam CF, et al. The 2019 International Society of Urological Pathology (ISUP) Consensus Conference on Grading of Prostatic Carcinoma. The American Journal of Surgical Pathology. 2020; 44: e87–e99. https://doi.org/10.1097/PAS.0000000000001497