Development and validation of machine learning models for predicting muscle-invasive bladder cancer: Comparative performance analysis with VI-RADS

Batu Akalin, Akif Erbin, Onuralp Nisli, Halil Lutfi Canat
Department of Urology, Health Science University, Basaksehir Cam Sakura City Hospital, Istanbul, Turkey

Cite as: Akalin B, Erbin A, Nisli O, et al. Development and validation of machine learning models for predicting muscle-invasive bladder cancer: Comparative performance analysis with VI-RADS. *Can Urol Assoc J* 2026 November 20; Epub ahead of print http://dx.doi.org/10.5489/cuaj.9728

Published online November 20, 2026

Corresponding author: Dr. Akif Erbin, Department of Urology, Health Science University, Basaksehir Cam Sakura City Hospital, Istanbul, Turkey; akiferbin@hotmail.com

***

ABSTRACT

Introduction: We aimed to develop and validate machine learning (ML)-based predictive models for the preoperative differentiation of muscle-invasive bladder cancer (MIBC) from non-muscle-invasive bladder cancer (NMIBC) using routinely available clinical and laboratory data, and to compare their diagnostic performance with the multiparametric magnetic resonance imaging (mpMRI)-based Vesical Imaging-Reporting and Data System (VI-RADS).
Methods: A total of 26 demographic, clinical, and biochemical variables were used to develop supervised ML models. Logistic regression, random forest, support vector machine, XGBoost, and LightGBM algorithms were trained to predict pathologic muscle invasion. The dataset was randomly divided into training (80%) and independent test (20%) cohorts. Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC). The performance of ML models was compared with a VI-RADS-based logistic regression model using the DeLong test, and model calibration was assessed using Brier scores.
Results: A total of 428 patients were included, comprising 317 with NMIBC and 111 with MIBC. In the independent test cohort, both the VI-RADS–based logistic regression model and the LightGBM model demonstrated high discriminative performance, with AUC values of approximately 0.93. Among ML methods, LightGBM showed the best overall performance. SMOTE improved sensitivity in most models, while calibration remained acceptable. DeLong analysis showed no significant difference between the LightGBM and VI-RADS models (p=0.49).
Conclusions: ML models based on routinely available clinical and laboratory variables achieved high diagnostic accuracy for predicting MIBC. The LightGBM model demonstrated

© 2026 Canadian Urological Association

performance comparable to the VI-RADS-based approach in the independent validation cohort and may serve as a complementary tool for risk stratification, particularly in settings where access to mpMRI is limited.

INTRODUCTION

Bladder cancer is a heterogeneous disease associated with significant morbidity and mortality worldwide. The precise preoperative assessment of muscle invasion is critical for determining suitable treatment strategies (1). The multiparametric magnetic resonance imaging (mpMRI)-based Vesical Imaging-Reporting and Data System (VI-RADS) has attained a degree of standardization; however, its application is predominantly confined to large, well-equipped facilities due to its substantial cost and the necessity for assessment by an experienced radiologist (2).

The increasing availability of real-world clinical data has spurred interest in computational techniques capable of uncovering predictive patterns that are not readily identifiable through traditional evaluation methods. Clinical decision support systems based on machine learning (ML) have shown effective performance in multiple oncologic areas by incorporating diverse patient-related variables to produce reliable predictive results (3). It has been shown that machine learning methods applied to routine clinical and laboratory data can significantly increase diagnostic accuracy and prognostic predictive power (4).

We aim to evaluate the capability of ML-based predictive models to estimate bladder cancer stage in a manner that is more cost-effective and accessible than the use of VI-RADS. In this context, we will incorporate patient factors potentially associated with bladder cancer stage, along with routinely collected clinical and biochemical parameters obtained during diagnostic work-up and preoperative evaluation, into the analysis. Should these models demonstrate satisfactory diagnostic performance, subsequent external validation in independent cohorts may support their adoption as a complementary—or potentially alternative—tool for preoperative staging, particularly in clinical settings where multiparametric MRI is unavailable.

METHODS

Ethical standards

This study was conducted at a tertiary referral urology center (Health Science University, Basaksehir Cam Sakura City Hospital, Department of Urology) following institutional ethics approval (Basaksehir Cam Sakura City Hospital, Scientific Research Ethics Committee, decision no: 2026-43 ; date: 24.02.2026).

Study design

In this analytical cross-sectional study, patients who underwent primary transurethral resection of the bladder tumor (TUR-BT) were identified retrospectively from hospital's digital medical record system (n = 1016). Among these, only patients who had preoperative

© 2026 Canadian Urological Association

multiparametric magnetic resonance imaging (mpMRI) with a recorded VI-RADS and were histopathologically verified to have urothelial carcinoma were included in the study (n = 504). Exclusion criteria were a history of TURBT performed for recurrent tumors, histopathological diagnoses revealing benign lesions or metastases originating from other primary malignancies, the absence of preoperative MRI performed in accordance with the VI-RADS protocol, and more than 10% missing data (n = 428) (Figure-1).

For each patient, 12 clinical (age and sex, tumor size, VI-RADS score, history of smoking, history of macroscopic hematuria, comorbidities such as hypertension, diabetes and coronary artery disease, use of antiplatelet medications, history of chemical exposure and family history of cancer) and 14 biochemical variables (hemoglobin level, white blood cell; WBC and platelet counts, serum creatinine and glucose levels, electrolyte measurements including sodium and potassium, coagulation parameters such as activated partial thromboplastin time; aPTT and international normalized ratio; INR, liver function tests including alanine aminotransferase; ALT and aspartate aminotransferase; AST, and urinalysis findings comprising erythrocyte count, leukocyte count, and urine specific gravity) , in conjunction with histological results and were retrieved to develop predictive models for pathological staging, classified as non–muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). Tumor size was determined based on preoperative ultrasound reports, cystoscopic findings, and operative reports, rather than MRI reports. Examinations of mpMRI were interpreted by multiple radiologists with dedicated experience in genitourinary imaging in our centre. Although centralized re-evaluation was not performed, all radiologists had substantial expertise in uro-oncologic imaging.

Machine learning models

A comprehensive set of machine learning models was developed to predict pathological upstaging. In order to make comparisons to a baseline logistic regression model was constructed using the VI-RADS score alone. Then multiple supervised learning algorithms were implemented to explore both linear and nonlinear decision boundaries. These included decision tree, random forest, naive Bayes, support vector machine, k-nearest neighbors, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost) classifiers. The dataset was randomly split into training and testing subsets in an 80:20 ratio, and all models were trained using the same predefined feature sets to ensure fair and transparent comparison across algorithms.

Data analysis

Given that 111 patients had muscle-invasive disease, the number of predictors included in each model was restricted to a maximum of ten variables in accordance with the event-per-variable principle to reduce the risk of overfitting. Feature selection for multivariable machine learning models was guided by mutual information analysis, and the ten most informative variables were retained while excluding the VI-RADS score where appropriate. All models were trained using these same predefined set of ten variables. Data preprocessing steps, including missing value imputation and categorical variable encoding, were integrated into model-specific pipelines. Class imbalance was addressed using the standard Synthetic

© 2026 Canadian Urological Association

Minority Over-sampling Technique (SMOTE) based on k-nearest neighbors (k=5). Oversampling was performed using the default sampling strategy, which balances the minority class to match the majority class within the training dataset. SMOTE was incorporated within the training process using a pipeline-based approach, ensuring that resampling was applied exclusively to the training data. For tree-based ensemble methods and kernel-based algorithms, hyperparameter optimization was performed using grid search with five-fold cross-validation.

Sample size

This study was designed as a retrospective analysis; therefore, no a priori sample size or power calculation was performed. To more accurately reflect the real-world patient population, all consecutive patients who met the predefined inclusion criteria and had complete clinical, radiological, and pathological data during the study period were included in the final analysis. The sample size was thus determined by the total number of eligible patients available in the institutional database.

Statistical analysis

All statistical analyses were performed using Python (version 3.11). Data handling and preprocessing were conducted using the pandas and NumPy libraries. Statistical testing was implemented using SciPy and statsmodels, while graphical outputs were generated with Matplotlib. Continuous variables were assessed for normality using the Kolmogorov–Smirnov test. Normally distributed variables were summarized as mean ± standard deviation and compared using the independent-samples t-test, whereas non-normally distributed variables were presented as median with interquartile range (IQR) and compared using the Mann–Whitney U test. Categorical variables were expressed as frequencies and percentages and compared using the chi-square test or Fisher's exact test, as appropriate.

Model performance was evaluated on an independent test set. Discriminative ability was assessed by plotting receiver operating characteristic (ROC) curves and calculating the area under the ROC curve (AUC) for each model. In addition, classification performance was quantified using the F1 score, while overall probabilistic accuracy was assessed using the Brier score. To further assess model stability and reduce variability associated with a single train–test split, repeated stratified cross-validation was additionally performed for all models. Model calibration was examined using calibration curves, along with calibration slope and intercept estimates. To further improve probability estimates, model recalibration was performed using Platt scaling with five-fold cross-validation applied exclusively to the training data. To formally compare discriminative performance between models, the DeLong test was applied to assess differences in AUC between the best-performing model—namely, the LightGBM model and the logistic regression model constructed using the VI-RADS score alone.

In order to address clinical applicability in diagnostically indeterminate cases, an additional subgroup analysis was performed in patients with a VI-RADS score of 3. Model performance within this subgroup was evaluated separately on the held-out test set using the same performance metrics.

© 2026 Canadian Urological Association

Aiming to improve the interpretability, Shapley Additive Explanations (SHAP) analysis was performed for the LightGBM model. Global feature importance was calculated using mean absolute SHAP values, and feature effects were further explored using SHAP summary plots. In addition, an error analysis was conducted to investigate the characteristics of misclassified cases. Patients with false-negative predictions were compared with correctly classified muscle-invasive cases.

All statistical tests were two-sided, and a p-value <0.05 was considered statistically significant.

RESULTS

Patients with MIBC were significantly older and had larger tumors compared with those with NMIBC. Hemoglobin levels were significantly lower, whereas platelet counts, serum creatinine levels, and urinary erythrocyte and leukocyte counts were significantly higher in the MIBC group. Serum sodium levels were slightly but significantly lower in patients with MIBC.

With respect to clinical variables, the distribution of VI-RADS scores differed markedly between groups, with higher scores being significantly more frequent in the MIBC cohort. Macroscopic hematuria was also significantly more common in patients with MIBC. No significant differences were observed between groups regarding sex distribution, diabetes mellitus, hypertension, coronary artery disease, antiplatelet use, family history of cancer, chemical exposure, smoking status, white blood cell count, glucose level, potassium level, liver enzymes, coagulation parameters, or urine specific gravity (Table-1).

The VI-RADS score demonstrated the highest mutual information value, followed by tumor size, urinary erythrocyte and leukocyte counts, and hemoglobin level. Among laboratory parameters, hemoglobin, coagulation indices, electrolyte levels, and urine-specific findings showed varying degrees of information contribution to the prediction of muscle invasion. Demographic and comorbidity-related variables, including age, sex, hypertension, diabetes mellitus, and family history of cancer, exhibited lower but non-zero mutual information scores.

LightGBM and the VI-RADS–based logistic regression model demonstrated the highest discriminative performance, with AUC values of approximately 0.93.

Among the machine learning–based approaches, the LightGBM model demonstrated the most favorable overall performance, achieving the highest AUC and a balanced profile across discrimination, classification, and calibration metrics, particularly when trained with SMOTE. The application of SMOTE was associated with improvements in sensitivity-related performance metrics across most machine learning models, while calibration performance, as reflected by Brier scores, remained within acceptable ranges for the top-performing algorithms (Table-3) (Figure-2).

To test overfitting in the LightGBM model training set performance metrics were also evaluated. The model achieved an AUC of 0.96 in the training cohort and 0.92 in the independent test cohort, indicating only a modest reduction in performance between datasets (Figure-3).

© 2026 Canadian Urological Association

Figure-4 illustrates that probability recalibration improved the agreement between predicted and observed probabilities for both the logistic regression and LightGBM models. Post-recalibration curves demonstrated closer alignment with the ideal diagonal line, indicating enhanced calibration performance.

There was no statistically significant difference in AUC between the LightGBM model and the VI-RADS–based logistic regression model (DeLong $Z = -0.69$, $p = 0.49$), indicating comparable discriminative performance (Table-4).

To further evaluate clinical applicability in diagnostically indeterminate cases, a subgroup analysis was performed in patients with a VI-RADS score of 3 ($n = 18$). Within this subgroup, the LightGBM model achieved high discriminative performance (AUC = 0.93), correctly identifying all muscle-invasive cases with no false negatives and only one false positive prediction.

SHAP analysis was performed for the LightGBM model for explanatory analysis. Mean absolute SHAP values were calculated and identified tumor size, urinary erythrocyte count and age as the most influential features driving model predictions (Table-5). To show magnitude and direction of feature effects a SHAP summary plot was illustrated (Figure-5). Consistent with these findings, error analysis revealed that false-negative cases were characterized by smaller tumor size and less pronounced urinary abnormalities compared with correctly identified muscle-invasive cases, suggesting that the model may underperform in patients with more subtle clinical presentations (Table-6).

DISCUSSION

This study involved the development and validation of several ML models utilizing commonly accessible clinical and laboratory characteristics to predict MIBC, and we compared their diagnostic efficacy with the mpMRI-based VI-RADS scoring system. Our results indicate that the LightGBM model attained superior discriminative performance, akin to that of the VI-RADS–based logistic regression model, with no statistically significant difference in AUC between the two methodologies. The implementation of class-balancing and probability recalibration strategies enhanced model performance and calibration, highlighting the viability of ML-based risk stratification utilizing non-imaging data. The findings indicate that data-driven models based on frequently gathered clinical data could act as an adjunct to VI-RADS, especially in environments with restricted access to high-quality mpMRI or expert radiological analysis.

Since Panebianco et al. developed the VI-RADS scoring system it has emerged as a highly reliable and standardized imaging-based tool for preoperative assessment of muscle invasion in bladder cancer, demonstrating consistently high diagnostic accuracy across multiple studies and meta-analyses(2,5,6,7). In a recent meta analysis Jazayeri et al. reported that the VI-RADS scoring system achieved excellent discriminative performance for the detection of muscle-invasive bladder cancer, with an AUC of 0.93 (95% CI: 0.91–0.95), which is consistent with the findings of the present study(2). Nevertheless, due to acquisition of high-quality multiparametric MRI is associated with substantial financial costs, requires advanced imaging infrastructure, and depends heavily on the availability of experienced uro-oncologic radiologists the widespread implementation of VI-RADS remains limited (8,9).

© 2026 Canadian Urological Association

Furthermore, interobserver variability and limited access to standardized mpMRI protocols may restrict its applicability in smaller centers or resource-limited healthcare settings (10,11). In contrast, machine learning–based approaches offer a more accessible and scalable alternative by leveraging routinely available clinical and laboratory data that are already obtained as part of standard diagnostic workflows (12,13). Such data-driven models can be implemented without specialized imaging equipment or advanced radiologic expertise, thereby potentially broadening access to accurate preoperative risk stratification. In this context, our findings underscore that ML-based models—particularly the LightGBM algorithm—can achieve discriminative performance comparable to that of a VI-RADS–only logistic regression model. Taken together, these results suggest that ML-based prediction using routine clinical data may represent a clinically meaningful adjunct, and potentially an alternative, to VI-RADS in selected settings.

To our knowledge, there is no published study that mirrors our analytical framework in a fully comparable manner by simultaneously evaluating multiple machine learning algorithms, assessing both discrimination and calibration, and directly comparing ML-based predictions with a VI-RADS–only model using formal statistical testing. Previous investigations on this topic have primarily focused on imaging-based prediction models (13,14,15). In contrast, our study adopts a more comprehensive and methodologically rigorous approach. By evaluating multiple complementary performance metrics—including AUC, F1 score, and Brier score—and employing DeLong's test to compare correlated ROC curves, we provide a multidimensional evaluation of predictive performance that extends beyond discrimination alone. Such a framework has been infrequently applied in the existing literature on bladder cancer staging. Moreover, the inclusion of a broad spectrum of demographic, clinical, laboratory, and urinalysis parameters likely enhanced the model's ability to capture subtle but clinically relevant patterns associated with muscle invasion. These methodological strengths, together with the richness of the analyzed variables, may explain the relatively high AUC values observed in our models.

Another key finding of this study is the performance of machine learning models in diagnostically indeterminate cases. In the subgroup of patients with a VI-RADS score of 3, the LightGBM model achieved excellent accuracy and successfully identified all muscle-invasive cases. This might suggest that machine learning approaches may provide meaningful clinical value in borderline cases, where imaging-based assessment alone is not helpful and treatment decisions are most challenging. It must be noted tho the sample size in this analysis was small and warrants cautious interpretation of these findings.

This study possesses multiple limitations that warrant acknowledgment. The retrospective and single-center approach may induce selection bias and restrict the generalizability of the findings to wider patient populations and diverse therapeutic environments. There was a moderate imbalance in our dataset for which we implemented SMOTE. Although it improves model sensitivity by balancing class distributions, it may introduce synthetic patterns that do not fully represent real-world data, potentially affecting generalizability. Secondly, while internal validation techniques such as cross-validation, bootstrap resampling, and the utilization of an independent test set were applied to strengthen

© 2026 Canadian Urological Association

model resilience, external validation in independent, multicenter cohorts is absent and necessary prior to clinical application. Thirdly, the radiologic evaluation relied on the reported VI-RADS ratings instead of a centralized reassessment by a single expert radiologist, potentially introducing interobserver variability. Nevertheless, the radiology department of our institution has specialized expertise in uro-oncologic imaging and possesses extensive experience in this field.

This study possesses several limitations that warrant acknowledgment. First, the retrospective and single-center design may introduce selection bias and limit the generalizability of the findings to broader patient populations and diverse clinical settings. There was a moderate class imbalance in the dataset, for which SMOTE was implemented. Although this approach improves model sensitivity by balancing class distributions, it may introduce synthetic patterns that do not fully reflect real-world data, potentially affecting generalizability. Second, although internal validation techniques—including cross-validation, bootstrap resampling, and the use of an independent test set—were employed to enhance model robustness, external validation in independent, multicenter cohorts is lacking and remains necessary prior to clinical application. Third, the use of TURBT pathology as the reference standard represents an inherent limitation, as understaging of T1 disease is a possible issue in bladder cancer. In our institution, repeat TURBT is routinely performed for patients with T1 disease, and in this study, all available pathological data were reviewed, with the highest pathological stage accepted as the final diagnosis. In patients who subsequently underwent radical cystectomy, the cystectomy specimen pathology was used as the reference standard. Nevertheless, despite these measures, the possibility of residual misclassification cannot be completely excluded. Finally, radiologic evaluation relied on reported VI-RADS scores rather than centralized reassessment by a single expert radiologist, which may introduce interobserver variability. However, the radiology department at our institution has substantial expertise in uro-oncologic imaging, which may partially mitigate this limitation.

Although the clinical and laboratory variables included are routinely accessible and augment the model's practicality, other potentially informative factors—such as molecular biomarkers, radiomic features, or longitudinal data—were excluded and could enhance predictive performance. In future research hybrid models can be created combining VI-RADS and clinical models and may be particularly valuable in improving risk stratification in borderline or indeterminate cases, where neither imaging nor clinical parameters alone provide sufficient diagnostic certainty. Furthermore, hybrid frameworks may facilitate more personalized clinical decision-making by integrating complementary sources of information.

The strengths of this study include the use of a comprehensive set of routinely accessible clinical and laboratory parameters, which enhances the practical applicability of the suggested models in real-world clinical settings. A direct comparison of several ML algorithms with the standard VI-RADS-based approach, utilizing strong statistical approaches such as DeLong testing, offers a clear and methodologically rigorous performance evaluation. Finally, the inclusion of histopathologically confirmed outcomes as the reference standard ensures the reliability of the predictive assessments.

© 2026 Canadian Urological Association

CONCLUSIONS

This study revealed that ML models, utilizing commonly accessible clinical and laboratory information, had robust diagnostic efficacy for the preoperative prediction of muscle-invasive bladder cancer. Among the assessed methodologies, the LightGBM model had the most advantageous overall performance and attained discriminative capability akin to that of the VI-RADS–based logistic regression model. These findings suggest that data-driven models based on routinely gathered clinical information may act as an adjunct to conventional imaging-based staging systems. With additional external validation, such methodologies could enhance clinical decision-making and refine risk categorization, especially in environments where access to multiparametric MRI or specialized radiologic expertise is limited.

DRAFT

© 2026 Canadian Urological Association

REFERENCES

1. European Association of Urology. EAU guidelines on muscle-invasive and metastatic bladder cancer 2025. Arnhem: European Association of Urology; 2025.
2. Jazayeri SB, Dehghanbanadaki H, Hosseini M, et al. Diagnostic accuracy of vesical imaging-reporting and data system (VI-RADS) in suspected muscle-invasive bladder cancer: A systematic review and diagnostic meta-analysis. *Urol Oncol* 2022;40:45-55. https://doi.org/10.1016/j.urolonc.2021.11.008
3. Oehring R, Ramasetti N, Ng S, et al. Use and accuracy of decision support systems using artificial intelligence for tumor diseases: A systematic review and meta-analysis. *Front Oncol* 2023;13:1224347. https://doi.org/10.3389/fonc.2023.1224347
4. Liu C, Liao Y, Wang D, et al. Machine learning for classification in lung cancer using routine clinical and laboratory data. *Ann Surg Oncol* 2025. https://doi.org/10.1245/s10434-025-18747-y
5. Panebianco V, Narumi Y, Altun E, et al. Multiparametric magnetic resonance imaging for bladder cancer: Development of VI-RADS (Vesical Imaging-Reporting and Data System). *Eur Urol* 2018;74:294-306. https://doi.org/10.1016/j.eururo.2018.04.029
6. Nesiu A, Novacescu D, Latcu S, et al. Diagnostic performance and interobserver agreement of the vesical imaging-reporting and data system (VI-RADS) in bladder cancer staging: A systematic review. *Medicina (Kaunas)* 2025;61:469. https://doi.org/10.3390/medicina61030469
7. Wang H, Luo C, Zhang F, et al. Multiparametric MRI for bladder cancer: Validation of VI-RADS for the detection of detrusor muscle invasion. *Radiology* 2019;291:668-74. https://doi.org/10.1148/radiol.2019182506
8. Pecoraro M, Cipollari S, Messina E, et al. Multiparametric MRI for bladder cancer: A practical approach to the clinical application of VI-RADS. *Radiology* 2025;314:e233459. https://doi.org/10.1148/radiol.233459
9. Dehghanpour A, Davenport MS, Muglia VF, et al. Vesical imaging reporting and data system: Overcoming barriers to global implementation. *AJR Am J Roentgenol* 2025. https://doi.org/10.2214/AJR.25.33634
10. Barchetti G, Simone G, Ceravolo I, et al. Multiparametric MRI of the bladder: Inter-observer agreement and accuracy with the vesical imaging-reporting and data system (VI-RADS) at a single reference center. *Eur Radiol* 2019;29:5498-506. https://doi.org/10.1007/s00330-019-06117-8
11. Laert J, Manso LN, de Santana TF, et al. Pitfalls and limitations of the vesical imaging-reporting and data system (VI-RADS): Challenges in the multiparametric magnetic resonance imaging assessment of bladder cancer. *Cureus* 2025;17:e99197. https://doi.org/10.7759/cureus.99197
12. Tsai IJ, Shen WC, Lee CL, et al. Machine learning in prediction of bladder cancer on clinical laboratory data. *Diagnostics (Basel)* 2022;12:203. https://doi.org/10.3390/diagnostics12010203
13. Ferro M, Falagario UG, Barone B, et al. Artificial intelligence in the advanced diagnosis of bladder cancer: Comprehensive literature review and future advancement. *Diagnostics (Basel)* 2023;13:2308. https://doi.org/10.3390/diagnostics13132308
14. Boca B, Caraiani C, Telecan T, et al. MRI-based radiomics in bladder cancer: A systematic review and radiomics quality score assessment. *Diagnostics (Basel)* 2023;13:2300. https://doi.org/10.3390/diagnostics13132300

© 2026 Canadian Urological Association

15. Huang X, Wang X, Lan X, et al. The role of radiomics with machine learning in the prediction of muscle-invasive bladder cancer: A mini review. *Front Oncol* 2022;12:990176. https://doi.org/10.3389/fonc.2022.990176

DRAFT

© 2026 Canadian Urological Association

FIGURES AND TABLES

Figure 1. Flowchart of the study.

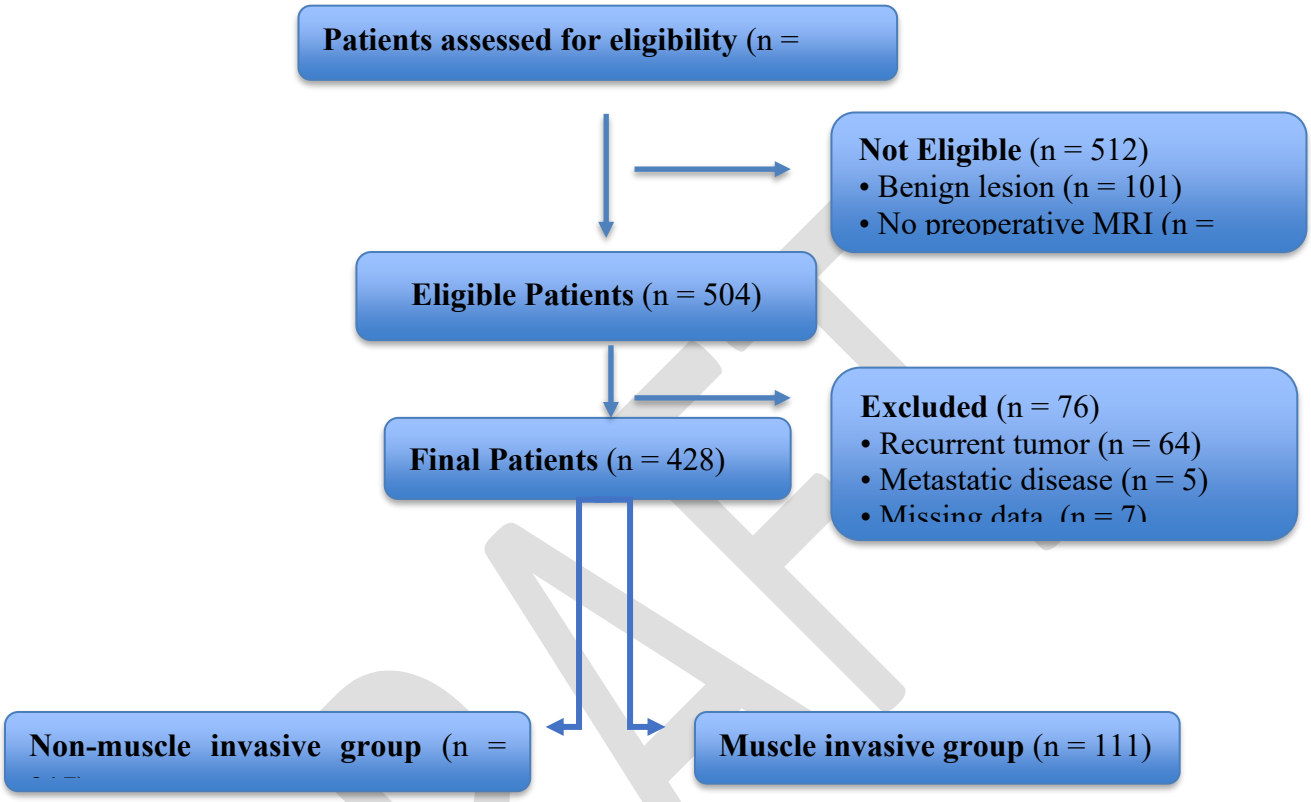
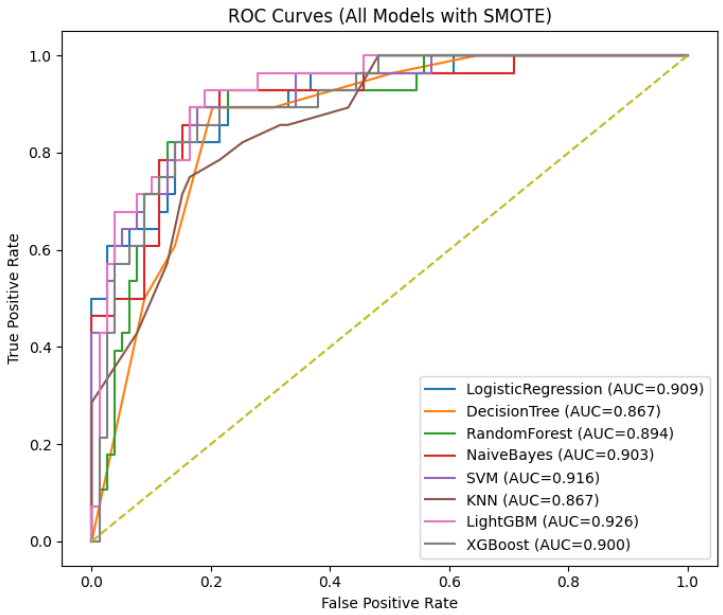


Figure 2. ROC curves of all machine learning models after application of the SMOTE for predicting MIBC.



© 2026 Canadian Urological Association

Figure 3. Receiver operating characteristic curves of the LightGBM model for the training and independent test cohorts.

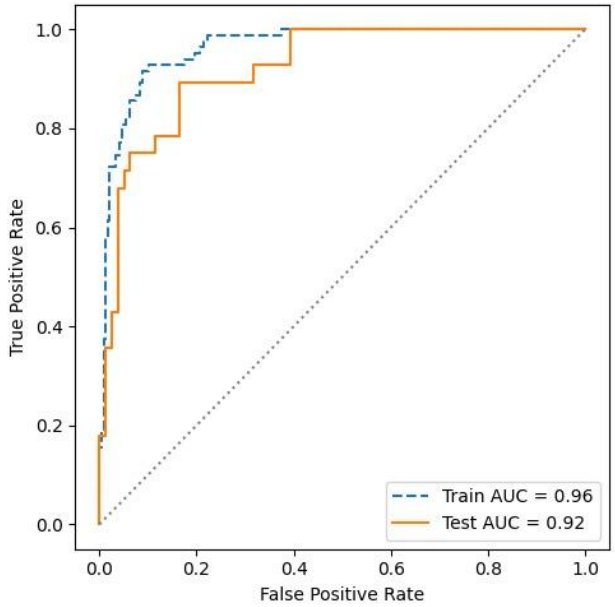
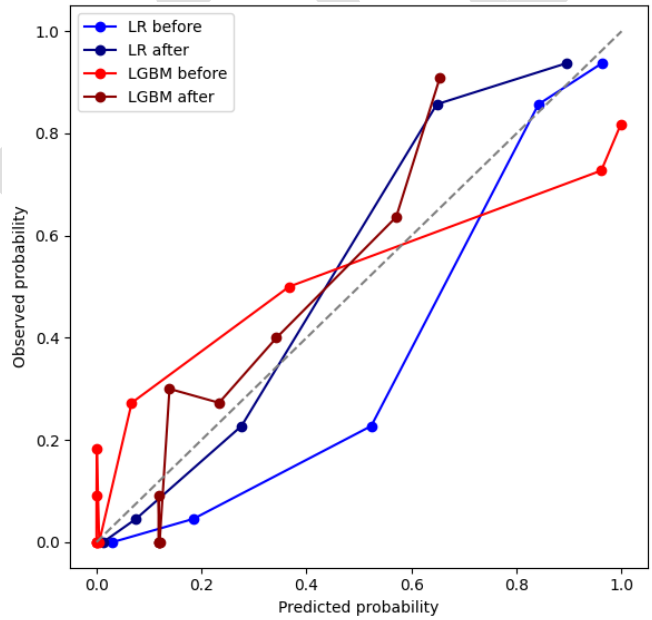


Figure 4. Calibration curves of the logistic regression and LightGBM models before and after probability recalibration.



© 2026 Canadian Urological Association

Figure 5. SHAP summary plot for LightGBM model.

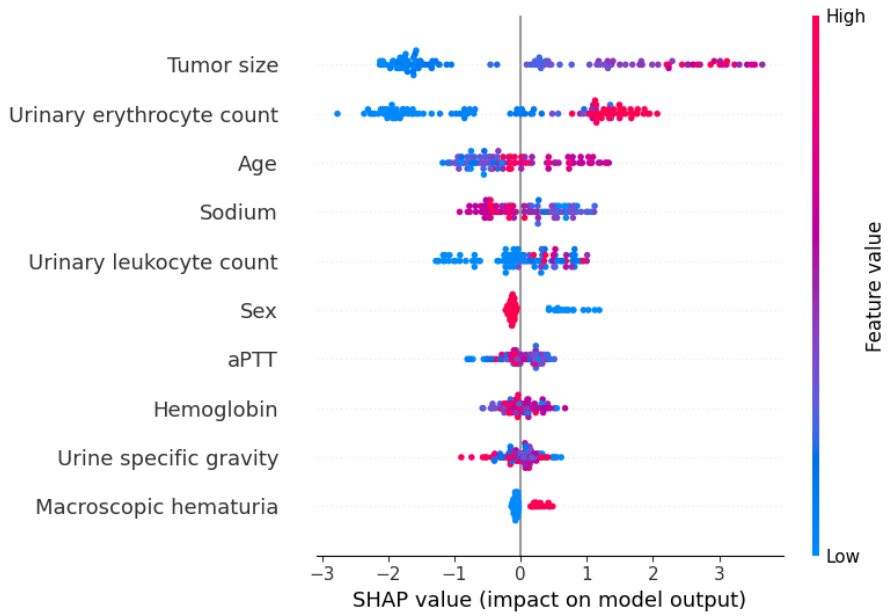


Table 1.Comparison of baseline demographic, clinical, laboratory, and VI-RADS characteristics between patients with NMIBC and MIBC			
Variable	**NMIBC (n=317)**	**MIBC (n=111)**	**p**
Age (years), mean ± SD	64.8±10.7	69.6±9.6	**<0.0001**[*]
Sex (male), n (%)	269 (84.9%)	88 (79.3%)	0.2257[§]
Tumor size (mm), median [IQR]	17.0 [11.0–29.0]	41.0 [32.0–54.5]	**<0.0001**[‡]
VI-RADS score, n (%)	No visible lesion: 33 (10.4%) 1: 40 (12.6%) 2: 151 (47.6%) 3: 82 (25.9%) 4: 8 (2.5%) 5: 3 (0.9%)	No visible lesion: 0 (0.0%) 1: 2 (1.8%) 2: 5 (4.5%) 3: 29 (26.1%) 4: 27 (24.3%) 5: 48 (43.2%)	**<0.0001**[§]
Smoking (pack-years), median [IQR]	30 [8.0–45.0]	30 [0–43.7]	0.223[‡]
History of macroscopic hematuria, n (%)	41 (12.9%)	48 (43.2%)	**<0.0001**[§]
Hypertension, n (%)	148 (46.7%)	63 (56.8%)	0.0862[§]
Diabetes mellitus, n (%)	85 (26.8%)	24 (21.6%)	0.3401[§]
Coronary artery disease, n (%)	83 (26.2%)	35 (31.5%)	0.3361[§]
Antiplatelet usage, n (%)	109 (34.4%)	45 (40.5%)	0.2946[§]
Chemical exposure, n (%)	17 (5.4%)	6 (5.4%)	1.0000[§]
Family history of cancer, n (%)	28 (8.8%)	14 (12.6%)	0.3337[§]
Hemoglobin (g/dL), mean ± SD	13.5±1.9	12.2±2.6	**<0.0001**[*]
White blood cell count (×10⁹/L), median [IQR]	8.10 [6.7–9.9]	8.40 [7.1–10.3]	0.0756[‡]

© 2026 Canadian Urological Association

Platelet count (×10⁹/L), mean ± **SD**	260.5 ± 77.7	295.3 ± 100.3	**0.0011**[*]
Serum creatinine (mg/dL), median [IQR]	0.9 [0.8–1.1]	1.1 [0.8–1.3]	**0.0004**[‡]
Serum glucose (mg/dL), median [IQR]	101.0 [90.0–127.0]	105.0 [92.0–128.5]	0.3344[‡]
Sodium (mmol/L), median [IQR]	141.0 [139.0–142.0]	139.0 [138.0–142.0]	**0.0049**[‡]
Potassium (mmol/L), median [IQR]	4.4 [4.2–4.7]	4.4 [4.2–4.7]	0.6082[‡]
aPTT (seconds), median [IQR]	29.3 [27.4–32.6]	29.8 [27.8–32.2]	0.3652[‡]
INR, median [IQR]	1.0 [0.9–1.1]	1.0 [0.9–1.1]	0.7264[‡]
ALT (U/L), median [IQR]	16.0 [12.0–21.0]	15.0 [9.5–19.0]	**0.0062**[‡]
AST (U/L), median [IQR]	17.0 [14.0–21.0]	16.0 [14.0–20.0]	0.2186[‡]
Urine RBC count, median [IQR]	27.0 [3.0–275.0]	409.0 [98.5–409.0]	**<0.0001**[‡]
Urine WBC count, median [IQR]	3.0 [2.0–13.0]	14.0 [4.5–41.0]	**<0.0001**[‡]
Urine specific gravity, mean ± **SD**	1016.8±6.8	1015.8±6.0	0.1442[*]

[*]Student's t-test. [‡]Mann–Whitney U test. [§]Chi-squared test. ALT: alanine aminotransferase; AST: aspartate aminotransferase; aPTT: activated partial thromboplastin time; INR: international normalized ratio; IQR: interquartile range; MIBC: muscle-invasive bladder cancer; NMIBC: non-muscle-invasive bladder cancer; RBC: red blood cell; SD: standard deviation; VI-RADS: Vesical Imaging-Reporting and Data System; WBC: white blood cell.

Table 2. Mutual information scores of variables used for feature selection in machine learning models

Feature	**MI score**
VI-RADS score	0.289
Tumor size	0.182
Urinalysis erythrocytes	0.091
Urinalysis leukocytes	0.090
Hemoglobin	0.070
History of macroscopic hematuria	0.047
aPTT	0.038
Sex	0.035
Age	0.034
Urine specific gravity	0.034
Sodium	0.026
Hypertension	0.022
Family history of cancer	0.015
Diabetes mellitus	0.013

aPTT: activated partial thromboplastin time; VI-RADS: Vesical Imaging-Reporting and Data System.

© 2026 Canadian Urological Association

Table 3. Diagnostic performance of machine learning models for predicting MIBC with and without SMOTE

Model	SMOTE						No SMOTE				
	AUC	Mean AUC	PPV	NPV	F1	Brier	AUC	PPV	NPV	F1	Brier
LightGBM	**0.926**	0.915	0.769	0.901	0.740	0.094	**0.916**	0.800	0.862	0.666	0.102
SVM	0.916	0.899	0.595	0.954	0.714	0.132	0.909	0.692	0.876	0.666	0.108
Logistic regression (VI-RADS only)	**0.933**	–	0.578	0.968	0.712	0.091	**0.932**	0.913	0.916	0.823	0.073
Logistic Regression	0.909	0.902	0.581	0.953	0.704	0.149	0.912	0.782	0.881	0.705	0.103
Naive Bayes	0.903	0.886	0.610	0.954	0.724	0.146	0.894	0.647	0.917	0.709	0.131
XGBoost	0.900	0.893	0.668	0.918	0.721	0.109	0.924	0.761	0.860	0.653	0.095
Random forest	0.894	0.882	0.677	0.931	0.741	0.121	0.913	0.772	0.870	0.680	0.107
Decision tree	0.867	0.842	0.556	0.951	0.684	0.135	0.866	0.692	0.876	0.666	0.121
KNN	0.867	0.863	0.535	0.921	0.647	0.161	0.861	0.705	0.822	0.533	0.126

Logistic regression (VI-RADS only) refers to a logistic regression model constructed using the VI-RADS score as the sole predictor. Model performance was evaluated with and without SMOTE to assess the impact of class imbalance handling. Higher AUC, PPV, NPV, and F1 values indicate better discriminative performance, whereas lower Brier scores indicate better calibration. AUC: area under the receiver operating characteristic curve; Brier: Brier score; F1: F1 score; NPV: negative predictive value; PPV: positive predictive value; SMOTE: synthetic minority over-sampling technique; VI-RADS: Vesical Imaging-Reporting and Data System.

Table 4. Comparison of discriminative performance between the LightGBM model and the VI-RADS–based logistic regression using the DeLong test

Model	AUC
LightGBM	0.926
Logistic regression (VI-RADS only)	0.933
DeLong Z-statistic	-0.69
DeLong p-value	0.49

AUC: area under the receiver operating characteristic curve; VI-RADS: Vesical Imaging-Reporting and Data System.

© 2026 Canadian Urological Association

Table 5. Global feature importance based on SHAP values in the LightGBM model	
Feature	**Mean absolute SHAP value**
Tumor size	1.68
Urinary erythrocyte count	1.36
Age	0.59
Sodium	0.48
Urinary leukocyte count	0.47
Sex	0.24
aPTT	0.22
Hemoglobin	0.20
Urine specific gravity	0.19
Macroscopic hematuria	0.12

aPTT: activated partial thromboplastin time.

Table 6. Error analysis for LightGBM model			
Variable	**True positive MIBC**	**False-negative MIBC**	**p**
Tumor size, mm, median [IQR]	57.0 [48.5–82.0]	40.0 [29.5–45.8]	**0.008**‡
Age, years, median [IQR]	71.0 [66.0–74.5]	65.0 [63.3–78.3]	0.389‡
Urinary erythrocyte count, median [IQR]	409.0 [55.8–409.0]	123.5 [24.3–409.0]	0.118‡
Urinary leukocyte count, median [IQR]	34.5 [14.3–48.8]	8.5 [3.0–40.5]	0.125‡
Sodium, mmol/L, median [IQR]	138.0 [137.5–140.3]	140.0 [138.8–143.3]	0.091‡
Hemoglobin, g/dL, median [IQR]	11.45 [9.23–12.72]	13.25 [11.32–13.80]	0.074‡
Macroscopic hematuria, n (%)	11 (68.8%)	3 (25.0%)	0.054§
Male sex, n (%)	13 (81.2%)	10 (83.3%)	1.000§

‡Mann-Whitney U test. §Chi-squared test. IQR: interquartile range.

DRAFT

© 2026 Canadian Urological Association