Research Square

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

Hybrid Machine Learning Models Integrating VI-RADS and Clinical Metrics for Bladder Cancer Staging

Batu Akalin
Department of Urology, Health Science University, Basaksehir Cam Sakura City Hospital
Akif Erbin
akiferbin@gmail.com
Department of Urology, Health Science University, Basaksehir Cam Sakura City Hospital
Merve Sam Ozdemir
Department of Radiology, Health Science University, Basaksehir Cam Sakura City Hospital
Halil Lutfi Canat
Department of Urology, Health Science University, Basaksehir Cam Sakura City Hospital

Research Article

Keywords: Bladder Neoplasms, Vesical Imaging-Reporting and Data System (VI-RADS), Machine Learning, Magnetic Resonance Imaging, Random Forest Model, Artificial Intelligence, Muscle Invasion

Posted Date: April 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9081158/v1

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

Additional Declarations: No competing interests reported.

Abstract

Objective

In bladder cancer, integrating imaging and non-imaging parameters may enhance diagnostic performance beyond the Vesical Imaging-Reporting and Data System (VI-RADS). This study aimed to develop and validate machine learning models incorporating VI-RADS scores with clinical and laboratory variables to predict muscle invasion and support individualized treatment decisions.

Materials and Methods

A total of 372 patients who underwent transurethral resection of bladder tumor between 2019 and 2024 and had preoperative mpMRI performed according to the VI-RADS protocol were retrospectively evaluated. VI-RADS scores were combined with demographic data, hematological indices, biochemical markers, and urinalysis findings to construct predictive models. Machine learning algorithms—including logistic regression, random forest, support vector machines, extreme gradient boosting, light gradient boosting machine, and deep neural networks—were developed and optimized. Model performance was assessed using receiver operating characteristic area under the curve (AUC), sensitivity, specificity, Brier score, and decision curve analysis (DCA) and compared with VI-RADS alone.

Results

Pathological muscle invasion (≥ T2) was identified in 103 (27.8%) of the 372 patients. VI-RADS alone yielded an AUC of 0.89. Models supported with clinical and laboratory parameters demonstrated significant improvement, particularly random forest (AUC = 0.95), support vector machines (AUC = 0.95), and logistic regression (AUC = 0.94). Calibration analysis of the isotonic regression–adjusted random forest model yielded a slope of 1.16 and an intercept of − 0.039, indicating probability estimates closely aligned with clinical reality. In DCA, the RF model outperformed both the “treat-all” and “treat-none” strategies, demonstrating clear net clinical benefit.

Conclusion

Integrating VI-RADS with clinical and laboratory parameters improves discrimination and calibration in predicting muscle invasion compared with imaging alone. The random forest model, in particular, may reduce misclassification at critical decision points—such as early radical cystectomy or neoadjuvant chemotherapy—and provide more reliable information for patient counseling.

Introduction

Bladder cancer is a common malignancy worldwide, with approximately 614,000 new cases diagnosed annually in both men and women. According to GLOBOCAN 2022, it ranks as the ninth most frequently diagnosed cancer globally (1). The age-standardized incidence rate (ASIR) is estimated at 5.6 per 100,000 for men and 1.9 for women. Despite advances in diagnosis and treatment, both the incidence and mortality of bladder cancer continue to rise globally (2). The development of bladder cancer is influenced by both environmental and genetic factors. Tobacco smoking remains the predominant risk factor, responsible for nearly half of all cases and increasing the risk approximately threefold compared to non-smokers (3). Occupational exposure to aromatic amines and polycyclic aromatic hydrocarbons, particularly in the dye, textile, and leather industries, is also a significant contributor (4).

Accurate staging plays a central role in determining prognosis and guiding management. The TNM classification system, developed by the American Joint Comitee on Cancer (AJCC) and Union Internationale Contre le Cancer (UICC), is the most widely adopted method for assessing the depth of bladder wall invasion, regional lymph node involvement, and distant metastasis (5). Tumors below stage T2 are classified as non–muscle-invasive bladder cancer (NMIBC), while T2 or higher tumors are considered muscle-invasive bladder cancer (MIBC). This distinction is clinically crucial, as NMIBC can often be managed with transurethral resection (TURBT), whereas MIBC typically requires radical cystectomy and/or multimodal treatment (4).

Cystoscopy remains the gold standard for diagnosis, offering direct visualization of bladder lesions (5). However, it is invasive, may cause patient discomfort, and provides limited information about the depth of invasion. Ultrasonography and computed tomography (CT) are useful for upper urinary tract evaluation but have limited accuracy in assessing the extent of muscle invasion (6). Multiparametric magnetic resonance imaging (mpMRI), standardized through the Vesical Imaging-Reporting and Data System (VI-RADS), has shown promising diagnostic accuracy in predicting muscle invasion (7). Nevertheless, interobserver variability and differences in implementation across centers remain significant limitations.

Accurate preoperative staging is critical for optimizing treatment strategy. Misclassification of tumor stage can result in unnecessary morbidity or suboptimal therapy, directly affecting patient outcomes. While VI-RADS has improved the radiologic assessment of bladder cancer, it relies solely on imaging data without incorporating clinical or biochemical variables that may also contribute to stage prediction.

This study hypothesizes that integrating imaging findings with clinical and biochemical data through machine learning algorithms can enhance the predictive performance for muscle invasion in bladder cancer. Therefore, we aimed to develop machine learning–based models combining VI-RADS scores with relevant clinical and biochemical parameters to predict muscle invasion preoperatively. Furthermore, we evaluated these models against VI-RADS alone and discussed their potential clinical applicability as preoperative decision-support tools.

Materials and Methods

Compliance with ethical standards

This research was conducted at the Urology Clinic of the University of Health Sciences Basaksehir Cam and Sakura City Hospital, a tertiary center in Istanbul, Turkey. This article presents a thesis study by Dr. Batu Akalin, a urology resident. The thesis adviser was Prof. Dr. Halil Lutfi Canat from the urology department.

The presented study was conducted in accordance with the principles of the Declaration of Helsinki (2013 revision) and Good Clinical Practice guidelines. Ethical approval was obtained from the Scientific Research Ethics Committee (Basaksehir Cam and Sakura City Hospital, decision No: 256; date: October 16, 2024). The study was retrospective, utilizing data acquired during clinical care without any supplementary interventions or sampling. Given that the study involved minimal risk and all personally identifiable information was removed, the committee granted a waiver of informed consent.

Study design

Data from 924 patients who underwent TURBT between November 2020 and June 2024 were retrospectively reviewed. Patients who underwent TURBT for recurrent tumors, those with histopathological diagnoses indicating benign lesions or metastases from other primary malignancies, and those who did not undergo preoperative MRI in accordance with the VI-RADS protocol were excluded from the study. After applying these exclusion criteria, a total of 372 patients were included in the final analysis (Figure-1).

A total of 8 clinical (patient's sex and age; history of smoking, macroscopic, and microscopic hematuria, hypertension, coronary artery disease, antiplatelet medication use, and chemical exposure) and 14 biochemical parameters (hemoglobin, white blood cell count (WBC), platelet count, serum creatinine, activated partial thromboplastin time (aPTT), international normalized ratio (INR), serum glucose, sodium, potassium, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and urinalysis findings including erythrocyte count, leukocyte count, and urine specific gravity), selected by the investigators, were collected along with each patient's pathological findings and VI-RADS scores. Complete urinalysis (CUA) parameters were obtained from urine samples provided at the time of the patients' initial outpatient evaluation, whereas biochemical parameters were derived from routine preoperative blood tests performed at the time the surgical decision was made.

Imaging and pathological assessments were conducted in a mutually blinded fashion. Radiologists were blinded to the histopathological outcomes, while pathologists performed their evaluations independently of the VI-RADS findings.

Pathological Assessment

The pathological evaluation was based on specimens obtained from the initial TURBT. In cases where a repeat TURBT (re-TURBT) was performed, the pathology report indicating the more advanced stage

between the two procedures was considered for analysis. All pathological assessments were performed and reported by the department of pathology at our institution. For the purposes of this study, pathological stages were categorized into two groups: <T2 (non-muscle-invasive) and ≥T2 (muscle-invasive) disease.

Multiparametric magnetic resonance imaging protocol

Magnetic resonance imaging (MRI) was performed in accordance with VI-RADS protocol using a 3.0 T scanner (Verio; Siemens, Erlangen, Germany). The multiplanar imaging capability of MRI minimized partial volume effects and optimized visualization for assessing the extent of muscular layer invasion. For contrast-enhanced T1-weighted sequences, a gadolinium-based, hydrophilic, nonionic macrocyclic contrast agent—gadobutrol (Gd-BT-DO3A; Gadovist, Bayer Schering Pharma AG, Berlin, Germany)—was administered as an intravenous bolus at a dose of 0.1 mL/kg using an MRI-compatible injector. To ensure adequate bladder distention, patients were instructed to drink 500–1000 mL of water approximately 30 minutes before imaging. All multiparametric MRI studies were interpreted by a single experienced urogenital radiologist.

Data Processing

Patients with more than 10% missing data were excluded from the study. For the remaining 16 patients with partially missing values, numerical variables were imputed using the median of the corresponding feature, whereas categorical variables were imputed using the mode. The dataset was randomly divided into a training set (80%) and a validation set (20%). In the training set, 214 cases were classified as <T2 and 82 cases as ≥T2 based on pathological staging. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data after the train–validation split.

Feature Selection and Machine Learning

Feature selection was performed using a Mutual Information (MI)–based filter approach implemented in Python (version 3.13). The relationship between each independent variable and the target variable (pathological outcome) was calculated individually using the *scikit-learn* library. The resulting MI scores were ranked in descending order, and the nine most informative variables were selected for model construction.

Subsequently, multiple supervised learning algorithms were developed, including logistic regression (LogReg), decision tree (DT), random forest (RF), Naïve Bayes, support vector machine (SVM), k-nearest neighbors (KNN), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost). These models were implemented using the *NumPy*, *Pandas*, *scikit-learn*, *LightGBM*, and *XGBoost* libraries. In addition, a deep neural network (DNN) model was constructed using the *TensorFlow* framework.

All models were tuned using a fivefold cross-validation approach implemented through the *GridSearchCV* function in *scikit-learn*. The model exhibiting the highest predictive power was considered optimal for subsequent analyses.

For the DT model, tree depth was varied from 1 to 10, and both *Gini* and *entropy* criteria were evaluated. For RF, the number of trees ranged from 300 to 900, with *Gini* as the splitting criterion. Additional parameters explored included maximum depth (None, 10, 20, 30), minimum samples for split (2, 5, 10), minimum samples per leaf (1, 2, 4), and maximum features (*sqrt*, *log2*). For SVM, *linear* and *radial basis function (RBF)* kernels were assessed, with *C* ranging from 10^{-3} to 10^{3} and *gamma* tuned between *scale* and *auto*.

For KNN, the number of neighbors was varied between 3 and 21, with weighting schemes (*uniform*, *distance*) and distance metrics (*Minkowski*, p = 1 or 2, corresponding to Manhattan and Euclidean distances). For XGBoost, the learning rate (*eta*) ranged from 0.01 to 0.2, maximum depth values of 3, 5, and 7 were tested, and *subsample* ratios of 0.8 and 1.0 were evaluated. For LightGBM, the number of leaves (15, 31, 63), maximum depth (−1, 5, 10), and subsample ratios (0.8, 1.0) were examined.

The DNN architecture was optimized by varying the number of hidden layers (1–3), neurons per layer (32, 64, 128), dropout rates (0.0, 0.2, 0.4), learning rates (0.001, 0.0003), number of epochs (50, 100), and batch sizes (32, 64). The ReLU activation function was used for hidden layers and sigmoid for the output layer, with optimization performed using the Adam algorithm.

Statistical Analysis

Descriptive analyses were performed using the Python programming language (version 3.13). The Pandas, NumPy, and SciPy libraries were utilized for statistical analysis. Continuous variables were presented as mean ± standard deviation and median values, whereas categorical variables were expressed as counts and percentages. The normality of continuous variables was assessed using the Kolmogorov-Smirnov test. Continuous variables were evaluated with the independent-samples t test or the Mann–Whitney U test according to their distributional properties, whereas categorical variables were assessed using the chi-square test.

Subsequently, to establish a baseline for comparison with machine learning models, a logistic regression model was constructed using only VI-RADS scores and pathological outcomes. A receiver operating characteristic (ROC) curve was generated, and the corresponding area under the curve (AUC) was calculated. All the models were then calibrated with isotonic regression. Differences in predictive performance among models and the VI-RADS only model were evaluated using the bootstrap test and Delong's test. The model with the best predictive performance was selected by the researchers for subsequent analysis. A calibration curve was drawn to test its calibration, slope, and intercept values, and the Brier score was calculated. The Brier scores of the best predictive model and the VI-RADS–based logistic regression model were statistically compared using a paired bootstrap test.

Finally, the calibrated and uncalibrated versions of the model with the highest AUC were evaluated using decision curve analysis (DCA), with 95% confidence intervals presented. The maximum net benefit (NB), threshold probability–dependent net benefits, and net intervention reduction per 100 patients (NRI/100) were computed, and a Clinical Impact Curve was generated for the calibrated model. A p-value of <0.05 was considered statistically significant.

Results

Differences in demographic characteristics, clinical features, and laboratory parameters between the two groups are summarized in Tables 1 and 2. Only the patient age exhibited a normal distribution. Patient age, VI-RADS score, erythrocyte and leukocyte counts in urinalysis, history of macroscopic hematuria, as well as serum hemoglobin, creatinine, sodium, platelet, and ALT levels were found to be significantly associated with tumor stage.

Table-1: Continous Variables

Variable (unit)	Non–muscle invasive group (n = 268)	Muscle invasive group (n= 104)	p-value
Age (years)	64.66 ± 10.73	69.82 ± 9.88	<0.001†
Hemoglobin (g/dL)	13.80 [2.70]	12.80 [3.35]	<0.001‡
White Blood Cell (×10⁹/L)	8.18 [3.16]	8.40 [3.06]	0.122‡
Platelet (×10⁹/L)	258.00 [104.00]	279.00 [115.50]	0.002‡
Creatinine (mg/dL)	0.93 [0.29]	1.12 [0.52]	<0.001‡
Activated Partial Thromboplastin Time (s)	29.30 [5.30]	29.80 [4.35]	0.278‡
International Normalized Ratio (INR)	1.00 [0.09]	1.00 [0.07]	0.520‡
Erythrocyte count (×10⁶/µL)	28.00 [260.75]	409.00 [310.50]	<0.001‡
Urinary Leukocyte (/HPF)	4.00 [13.00]	16.00 [40.00]	<0.001‡
Urine Specific Gravity	1017.00 [10.25]	1015.00 [9.00]	0.058‡
Smoking (pack-years)	30.00 [45.50]	25.00 [45.00]	0.168‡
Glucose (mg/dL)	101.00 [38.50]	106.00 [39.50]	0.290‡
Sodium (mmol/L)	141.00 [3.00]	139.00 [3.00]	0.002‡
Potassium (mmol/L)	4.48 [0.53]	4.51 [0.52]	0.694‡
Alanine Aminotransferase (U/L)	16.00 [9.00]	15.00 [9.50]	0.005‡
Aspartate Aminotransferase (U/L)	17.00 [7.00]	16.00 [6.00]	0.251‡

†: Independent-samples *t*-test; *‡:* Mann–Whitney *U* test

Table-2: Categorical Variables

Variable	Category	Non–muscle invasive group (n = 268)	Muscle invasive group (n= 104)	p-value
VI-RADS score	No Visible Lesion	28 (10.4%)	0 (0.0%)	<0.001§
	1	37 (13.8%)	1 (1.0%)	
	2	122 (45.5%)	5 (4.9%)	
	3	73 (27.2%)	29 (28.2%)	
	4	5 (1.9%)	23 (22.3%)	
	5	3 (1.1%)	45 (43.7%)	
Macroscopic hematuria	Absent	232 (86.6%)	59 (57.3%)	<0.001§
	Present	36 (13.4%)	44 (42.7%)	
Hypertension (HT)	Absent	145 (54.1%)	44 (42.7%)	0.0645§
	Present	123 (45.9%)	59 (57.3%)	
Coronary artery disease (CAD)	Absent	202 (75.4%)	69 (67.0%)	0.1339§
	Present	66 (24.6%)	34 (33.0%)	
Antiplatelet use	No	181 (67.5%)	60 (58.3%)	0.1194§
	Yes	87 (32.5%)	43 (41.7%)	
	Yes	28 (10.4%)	14 (13.6%)	
Chemical exposure	No	251 (93.7%)	97 (94.2%)	1.0000§
	Yes	17 (6.3%)	6 (5.8%)	
Sex	Female	40 (14.9%)	22 (21.4%)	0.1828§
	Male	228 (85.1%)	81 (78.6%)	

§: Chi-square test

After splitting the dataset into training and validation subsets, SMOTE was applied to the training data to address class imbalance. To reduce data noise and improve model performance, a MI based filter approach was then used for feature ranking. The nine highest-ranked features were selected and used for model construction and validation (Table 3).

For baseline comparison, a logistic regression model was developed using only the VI-RADS score (Figure 2). The accuracy of this model was calculated as 85.3%. The VI-RADS-based model correctly

classified non–muscle-invasive cases (pathology < T2) with a recall of 96.3%, whereas the recall for muscle-invasive cases (pathology ≥ T2) was 57.1%. The F1 score was 68.6%. Although the overall sensitivity and specificity were balanced, the model showed lower sensitivity in predicting muscle invasion. The ROC curve was plotted, yielding an AUC of 0.89, and the corresponding confusion matrix was generated (Figure 3).

Table-3: Parameters Used in the Models and Their MI Scores

Feature	MI Score
VI-RADS	0.275
Erythrocyte (urinalysis)	0.090
Leukocyte (urinalysis)	0.067
Age	0.048
Platelet count	0.030
History of macroscopic hematuria	0.024
Urine specific gravity	0.018
White blood cell count (WBC)	0.017
Alanine aminotransferase (ALT)	0.014

Machine learning models were constructed and subsequently calibrated using isotonic regression. ROC curves were generated for both calibrated and uncalibrated models (Figures 4 and 5). Among the calibrated models, logistic regression achieved an accuracy of 87%, precision of 79%, F1 score of 75%, recall of 71%, and an AUC of 0.94 when trained with the top nine features. The decision tree model reached 84% accuracy, 85% precision, 65% F1 score, 52% recall, and an AUC of 0.91. The random forest model demonstrated the strongest overall performance, with 89% accuracy, 84% precision, 80% F1 score, 76.2% recall, and an AUC of 0.95. The support vector machine (SVM) achieved 83% accuracy, 67% precision, 71% F1 score, 76% recall, and an AUC of 0.95. The k-nearest neighbors (KNN) model yielded 81% accuracy, 65% precision, 68% F1 score, 71% recall, and an AUC of 0.90. The Naïve Bayes model achieved 80% accuracy, 64% precision, 65% F1 score, 67% recall, and an AUC of 0.93. The XGBoost model reached 85% accuracy, 73% precision, 74% F1 score, 76% recall, and an AUC of 0.92. The LightGBM model showed 87% accuracy, 79% precision, 75% F1 score, 71% recall, and an AUC of 0.92. The deep neural network (DNN) model achieved an accuracy of 86.7%, precision of 82.4%, F1 score of 73.7%, recall of 66.7%, specificity of 94.4%, and an AUC of 0.90 (Table 4). Calibration curves were plotted for all models, and their Brier scores were calculated.

Table-4: Characteristics of the Models

Model	AUC	Brier Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression (VI-RADS only)	0.887	0.1074	0.853	0.857	0.571	0.686
Support Vector Machine (SVM)	0.947	0.0927	0.827	0.667	0.762	0.711
Random Forest	0.946	0.0790	0.893	0.842	0.762	0.800
Logistic Regression	0.944	0.0832	0.867	0.789	0.714	0.750
Naïve Bayes	0.930	0.1034	0.800	0.636	0.667	0.651
XGBoost	0.918	0.1012	0.853	0.727	0.762	0.744
Decision Tree	0.906	0.0982	0.840	0.846	0.524	0.647
k-Nearest Neighbors (KNN)	0.903	0.1223	0.813	0.652	0.714	0.682
LightGBM	0.896	0.1062	0.867	0.789	0.714	0.750
Deep Neural Network (DNN)	0.900	0.1152	0.867	0.824	0.667	0.737

Among the calibrated models, the highest AUC value was observed in the Support Vector Machine (SVM) model (0.947), whereas the lowest belonged to the LightGBM model (0.896). Overall, the Random Forest model demonstrated superior performance, characterized by high accuracy, strong F1 score, high specificity, and balanced sensitivity, while also achieving the lowest Brier score among all models. The remaining algorithms exhibited moderate to high ranges of accuracy and F1 scores.

The predictive performances of all machine learning models were compared with that of the model constructed using only the VI-RADS score, employing both the Bootstrap and DeLong tests (Table 5). According to the Bootstrap analysis, the predictive performances of the Random Forest, SVM, and Logistic Regression models were statistically superior to the VI-RADS–only model, with p-values of 0.005, 0.030, and 0.0345, respectively. In contrast, the DeLong test indicated that only the Random Forest model demonstrated a statistically significant improvement in predictive performance compared with the VI-RADS–only model ($p = 0.030$).

Table-5: Comparison of the Models with the VI-RADS–Only Model

Model	Bootstrap p-value	DeLong p-value
Support Vector Machine	0.0300	0.075
Random Forest	0.0050	0.030
Logistic Regression	0.0345	0.091
Naïve Bayes	0.1220	0.274
XGBoost	0.1605	0.317
Decision Tree	0.2175	0.444
k-Nearest Neighbors	0.3445	0.702
LightGBM	0.3845	0.756
Deep Neural Network	0.1805	0.354

Precision–recall (Figure 6) and calibration curves (Figure 7) were generated for the Random Forest model, which exhibited the highest predictive performance. The calibration slope and intercept were calculated as 1.16 and −0.039, respectively. The Brier scores for the logistic regression model based solely on the VI-RADS score and for the calibrated Random Forest model were 0.107 and 0.079, corresponding to Brier Skill Scores (BSS) of 37.6 and 54, respectively. A paired bootstrap test was conducted to compare these two models, yielding a p-value of 0.027 (95% CI: −0.0573 to −0.0031), indicating a statistically significant difference.

For the same Random Forest model, DCA and clinical impact curves were plotted with 95% confidence intervals (Figure 8). The maximum net benefit of the model was calculated as 0.26. Based on different threshold probabilities, the corresponding net benefits and the net reduction in interventions per 100 patients (NAI/100) were calculated and summarized in Table 6.

Table-6: Net Benefit and NAI/100 Values of the Random Forest Model

Threshold Probability (p)	Net Benefit (Model)	Net Benefit (Treat All)	NAI / 100 Patients	Positive (%)	True Positive (%)
0.05	263	239	42.7	57.3	28.0
0.10	234	195	33.4	41.3	25.3
0.20	212	90	46.2	34.7	24.0
0.30	184	-46	50.9	32.0	22.7
0.40	184	-178	56.9	29.3	22.7
0.50	174	-416	61.1	22.7	20.0
0.60	134	-775	62.1	20.0	17.3

Discussion

Since its introduction by Panebianco et al. in 2018, the VI-RADS system has been increasingly utilized to predict the stage of bladder cancer (7). Initially, small-scale studies aimed at testing its accuracy and applicability were conducted; over time, these were followed by multicenter prospective studies evaluating interobserver agreement, sensitivity, and specificity of the system (8, 9). Different versions and adaptations of VI-RADS have subsequently emerged, demonstrating that the scoring system remains reliable despite variations in image quality, MRI protocols, and observer experience across centers (10). The system has now been incorporated into both European Association of Urology (EAU) and American Urological Association (AUA) guidelines (4, 11) and has become progressively integrated into routine clinical practice. One notable example of this integration is its use as an important reference point during multidisciplinary tumor board discussions of cases with suspected muscle invasion.

However, the increasing clinical adoption of VI-RADS has also prompted new research directions. The observer-dependent nature of radiologic scoring and the variability in image quality have highlighted the need for more objective and automated approaches. In this context, artificial intelligence and machine learning–based methods hold promise for enhancing diagnostic accuracy by supporting the interpretation of VI-RADS scores. Numerous studies in the literature have investigated the integration of VI-RADS with machine learning, most of which have focused on augmenting the score with radiomic features (12, 13).

Machine learning has shown substantial progress over the past decade across multiple disciplines. Applications have been developed using diverse data types, including genomic, transcriptomic, and proteomic data, imaging data, electronic health records, and clinical laboratory data (14, 15). One of the important applications of machine learning, as in the present study, is the improvement of diagnostic accuracy. Obermeyer et al. suggested that applying machine learning to clinical laboratory data could markedly enhance both prognostic and diagnostic precision (16). Accordingly, a decision-support system based on VI-RADS scores and clinical laboratory parameters may provide a rapid and cost-effective solution for improving diagnostic accuracy.

In our study, clinical laboratory data were integrated with machine learning algorithms to improve the preoperative prediction of bladder cancer stage. One of the primary challenges encountered was the imbalance of the dataset. An unequal distribution between classes is referred to as imbalanced data (17). Numerous studies have demonstrated that imbalanced datasets can cause classification problems during the training of machine learning algorithms (18). Similarly, in our dataset, the class imbalance was expected to adversely affect the classification performance of the models. To mitigate this issue, we applied SMOTE, one of the most widely used oversampling methods. After splitting the dataset into training and test subsets, SMOTE was applied exclusively to the training data to synthetically generate new samples for the minority class, thereby balancing the class distribution. This approach is widely

recognized to improve classification performance and enhance the generalizability of machine learning models (19).

Another challenge in this study was the presence of missing data in some patients. Patients without VI-RADS scores were excluded at the beginning of the study. Since the remaining variables were routinely collected during preoperative evaluation, missing data were limited to only 16 patients. To ensure data integrity and model reliability, patients with more than 10% missing data were excluded from the analysis. The literature reports several effective methods to address missing data, including collecting additional samples, excluding incomplete cases, imputing missing values using mean or median substitution, and applying data imputation techniques (20). In our study, missing continuous variables were imputed with the median value of the corresponding variable group, whereas missing categorical variables were replaced with the mode of their respective groups.

The associations between biochemical parameters and bladder cancer stage and prognosis have been investigated in numerous prior studies. Although anemia has not been consistently associated with stage, it has been linked to prognosis. Additionally, studies have reported that hematologic indices incorporating WBC and platelet counts correlate with higher tumor stage. Advanced-stage bladder cancer, due to its higher tumor burden, may induce a more pronounced inflammatory response, thereby affecting WBC and platelet levels more markedly than in early-stage disease (21, 22). The presence of macroscopic hematuria has also been reported to be associated with muscle invasion (23). In advanced-stage tumors, more prominent hematuria may lead to greater reductions in hemoglobin levels compared with early-stage disease, potentially explaining the observed correlation between lower hemoglobin and higher tumor stage. Similarly, advanced-stage tumors may contribute to alterations in creatinine and electrolyte levels through multiple mechanisms.

In our analysis, urinary parameters such as erythrocyte count, leukocyte count, and urine specific gravity —used in our models—were also found to be significantly associated with tumor stage. Previous studies have shown that pyuria may be related to higher stage and grade in bladder cancer (24). Furthermore, several studies have reported a positive correlation between patient age and bladder cancer stage (25). The emergence of these variables through mutual information–based feature selection demonstrates that the models rely not only on complex mathematical representations but also on clinically interpretable and pathophysiologically meaningful patterns. In contrast, comorbidities, smoking history, and chemical exposure were not found to be associated with tumor stage in our dataset and were therefore excluded from model development.

Ueno et al. reported, in a multicenter and multi-reader study, AUC values of 0.84 (range, 0.83–0.85) for inexperienced readers and 0.88 (range, 0.82–0.91) for experienced readers, with an overall mean AUC of 0.87 for all readers (8). In our study, the logistic regression model constructed using only the VI-RADS and pathology results yielded an AUC of 0.89, consistent with findings reported in the literature. The principal finding of our research, however, is that the inclusion of simple clinical and laboratory parameters alongside VI-RADS improved the discriminative ability and probability calibration for

predicting muscle invasion compared with an image-only approach. While the logistic regression model based solely on VI-RADS demonstrated high accuracy, it showed limited sensitivity in identifying ≥T2 cases. In contrast, calibrated methods such as RF, SVM, and logistic regression exhibited statistically superior AUC values according to the Bootstrap test, and the RF model demonstrated a significantly lower Brier score and higher BSS compared with the VI-RADS-only model. These findings suggest that easily obtainable indicators such as hematologic, biochemical, and clinical history parameters provide additional independent information beyond imaging-based risk and may contribute to clinical benefit.

The PROBAST+AI framework emphasizes the importance of evaluating model calibration and explicitly recommends its analysis (26). For our isotonic regression–calibrated RF model, the calibration slope and intercept were calculated as 1.16 and −0.039, respectively. Although the slope value was slightly greater than 1, its proximity to 1 indicates a mild tendency toward underfitting; the model showed a slight optimism at lower probability ranges but remained overall acceptably calibrated. We believe the clinical relevance of these calibration results deserves particular attention. The significantly lower Brier score of the isotonic-calibrated RF model compared with the VI-RADS-only logistic regression model indicates not only superior predictive accuracy but also that the probability estimates produced by the RF model are more consistent with clinical reality. Such calibration improvement may reduce the risk of overtreatment or delayed intervention near critical therapeutic thresholds, such as decisions regarding early radical cystectomy or neoadjuvant chemotherapy.

To further interpret the clinical utility of our model, the DCA results warrant discussion. The NB curve of the model remained above both the “treat all” and “treat none” strategies, demonstrating a clear clinical advantage. Alongside NB values calculated for different threshold probabilities, we also compared the proportions of predicted positive cases and true positives. When the threshold probability was set at 5%, corresponding to the highest calculated net benefit, treating patients with probabilities above this threshold as T2 cases would correctly identify all true positives but would result in 29.7 false positives per 100 patients. When the threshold was increased to 50%, only 2.7 false positives per 100 patients were observed, at the cost of missing 8 true positive cases. We suggest that this threshold selection should be guided by clinician experience and preference.

The integration of machine learning models into clinical practice should be evaluated not only in terms of accuracy and calibration but also within the framework of ethical responsibility. Data were anonymized, securely stored on encrypted institutional servers, and accessible only to authorized researchers. To minimize algorithmic bias, particular care was taken during both the study design and model development phases. Methodologically, one of the main internal validity strengths of this study was the independent and blinded reporting by radiologists and pathologists, ensuring that imaging data, including VI-RADS scores, did not influence pathology results. In addition, SMOTE was applied strictly within the training dataset, and imputation, scaling, and feature selection were performed within a cross-validation pipeline to prevent data leakage and optimism bias, which are common in imbalanced clinical datasets.

There are, however, several limitations to this study. It is a single-center, retrospective analysis. The VI-RADS score may be affected by protocols and observer variability across institutions. The use of a single reader and the lack of modeling for MRI protocol differences may limit the generalizability of our findings. Therefore, recalibration (e.g., using Platt scaling or isotonic regression) may be necessary before applying the model in external settings. Although SMOTE enhanced class separation by enriching class boundaries, it may have affected the probability distributions; therefore external validation is warranted. Another limitation is the sample size: the study included 372 patients, and the relatively modest performance of the DNN model may be attributable to insufficient sample volume.

Finally, this study demonstrates that the integration of low-cost, easily accessible laboratory parameters with VI-RADS can provide a scalable contribution to clinical decision support. Our findings suggest that enhanced preoperative stage prediction could enable more personalized decision-making in processes such as TUR-BT technique selection, early referral for neoadjuvant chemotherapy or radical cystectomy, and patient counseling. The next logical step is to evaluate the model's portability and recalibration needs through multicenter, multi-reader, time-based external validation and to assess the feasibility and clinical impact (in terms of time, cost, and patient outcomes) of integrating a simple electronic decision-support tool (VI-RADS plus selected parameters) into clinical workflow. Since our study focused on preoperative stage prediction using routinely available parameters, non-routine laboratory assays were excluded. Future studies may explore the inclusion of less conventional variables that could improve stage prediction, as well as the use of VI-RADS and additional parameters to predict prognosis or recurrence risk. Moreover, applying radiomics-based machine learning could further enhance model performance by allowing artificial intelligence to generate a combined imaging–clinical prediction with higher calibration and interpretability. Such a model could eventually be deployed as a simple online calculator or mobile application for clinical use.

In summary, our study demonstrated that incorporating clinical and laboratory signals into VI-RADS–based prediction provides significant improvements in both discrimination and calibration performance, while maintaining internal validity through a leakage-free workflow and blinded evaluation. With further external validation and decision-based analyses, this approach has the potential to evolve into a practical clinical decision-support tool for preoperative staging in bladder cancer.

Conclusion

Machine learning algorithms, particularly when integrated with multiple clinical and biochemical parameters, serve as robust tools for predicting muscle invasion in bladder tumors, demonstrating superior discriminatory and calibration performance compared with the use of the VI-RADS alone. If validated by multicenter prospective studies, VI-RADS–based hybrid machine learning models may serve as practical and scalable tools for clinical decision support. This approach could facilitate more precise preoperative risk stratification and improve multidisciplinary decision-making in bladder cancer management.

Declarations

Consent to Participate: Because this study was retrospective and involved no interventions beyond routine clinical care, the Institutional Review Board granted a waiver of informed consent. All patient data were anonymized prior to analysis.

Ethics Approval: This study was approved by the Scientific Research Ethics Committee of Başakşehir Çam and Sakura City Hospital (Approval No: 256; Date: October 16, 2024).

Funding: The authors received no financial support for the research, authorship, or publication of this article.

Clinical trial number: not applicable

References

1. GLOBOCAN 2022: Bladder Cancer 9th Most Common Worldwide. World Bladder Cancer Patient Coalition, 14 Feb. 2024. https://worldbladdercancer.org/news_events/globocan-2022-bladder-cancer-is-the-9th-most-commonly-diagnosed-worldwide/
2. Zhang Y, Rumgay H, Li M, Yu H, Pan H, Ni J. The global landscape of bladder cancer incidence and mortality in 2020 and projections to 2040. J Glob Health. 2023;13:04109. 10.7189/jogh.13.04109. PMID: 37712386; PMCID: PMC10502766.
3. Cumberbatch MG, Kwesi et al. Dec. Epidemiology of Bladder Cancer: A Systematic Review and Contemporary Update of Risk Factors in 2018. European Urology, vol. 74, no. 6, 2018, pp. 784–95. 10.1016/j.eururo.2018.09.001
4. European Association of Urology. EAU Guidelines on Muscle-Invasive and Metastatic Bladder Cancer, 2025. European Association of Urology; 2025.
5. European Association of Urology. EAU Guidelines on Non-muscle-invasive Bladder Cancer, 2025. European Association of Urology; 2025.
6. Lee C, Hung no et al. 6, June 2017, pp. 1193–205. 10.2214/ajr.16.17114
7. Panebianco V, Narumi Y, Altun E, Bochner BH, Efstathiou JA, Hafeez S, Huddart R, Kennish S, Lerner S, Montironi R, Muglia VF, Salomon G, Thomas S, Vargas HA, Witjes JA, Takeuchi M, Barentsz J, Catto JWF. Multiparametric Magnetic Resonance Imaging for Bladder Cancer: Development of VI-RADS (Vesical Imaging-Reporting And Data System). Eur Urol. 2018;74(3):294–306. Epub 2018 May 10. PMID: 29755006; PMCID: PMC6690492.
8. Ueno Y, et al. VI-RADS: Multiinstitutional Multireader Diagnostic Accuracy and Interobserver Agreement Study. Am J Roentgenol. May 2021;216(5):1257–66. 10.2214/ajr.20.23604.

9. Jazayeri SB 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. Urologic Oncology: Seminars Original Investigations, 40, 2, pp. 45–55, 10.1016/j.urolonc.2021.11.008
10. Pecoraro M et al. Mar. Multiparametric MRI for Bladder Cancer: A Practical Approach to the Clinical Application of VI-RADS. Radiology, 314, 3, 2025, 10.1148/radiol.233459
11. Holzbeierlein J et al. Diagnosis and Treatment of Non–Muscle Invasive Bladder Cancer: AUA/SUO Guideline—2024 Amendment. The Journal of Urology, 2024. American Urological Association.
12. Zheng Z, Xu F, Gu Z, Yan Y, Xu T, Liu S, Yao X, Combining Multiparametric. Score to Preoperatively Differentiate Muscle Invasion of Bladder Cancer. Front Oncol. 2021;11:619893. PMID: 34055600; PMCID: PMC8155615. MRI Radiomics Signature With the Vesical Imaging-Reporting and Data System (VI-RADS).
13. Wang W. Integrating Radiomics with the Vesical Imaging-Reporting and Data System to Predict Muscle Invasion of Bladder Cancer. Urologic Oncology: Seminars and Original Investigations, vol. 41, no. 6, p. 294
.e1-294.e8
, doi:10.1016/j.urolonc.2022.10.024.
14. Lin M, et al. Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice. J Transl Med. 2025;23:388. https://doi.org/10.1186/s12967-025-06425-2.
15. You J, et al. Advancing Laboratory Medicine Practice With Machine Learning: Swift yet Exact. Ann Lab Med. 2025;45(1):22–35. 10.3343/alm.2024.0354. Epub 2024 Nov 26. PMID: 39587856; PMCID: PMC11609717.
16. Obermeyer Z, Emanuel EJ. Predicting the Future-Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375:1216–9.
17. Kaur H, Pannu HS, Malhi AK. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput Surv. 2019;52:79.
18. Aubaidan BH, et al. A review of intelligent data analysis: Machine learning approaches for addressing class imbalance in healthcare - challenges and perspectives. Intell Data Anal. 2025;29(3):699–719. 10.1177/1088467X241305509.
19. Elreedy D. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. Information Sciences, 505, pp. 32–64, 10.1016/j.ins.2019.07.070
20. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep. 2018;8:663.
21. Chen C et al. Preoperative Anemia as a Simple Prognostic Factor in Patients with Urinary Bladder Cancer. Med Sci Monit. 2017;23:3528–3535. 10.12659/msm.902855. PMID: 28723884; PMCID: PMC5531533.

22. Li DX, et al. Prognostic value of preoperative neutrophil-to-lymphocyte ratio in histological variants of non-muscle-invasive bladder cancer. Investig Clin Urol. 2021;62(6):641–9. https://doi.org/10.4111/icu.20210278.
23. Jakus D, et al. The Impact of the Initial Clinical Presentation of Bladder Cancer on Histopathological and Morphological Tumor Characteristics. J Clin Med. 2023;12(13):4259. 10.3390/jcm12134259. PMID: 37445294; PMCID: PMC10342402.
24. Poletajew Set al, et al. Preoperative pyuria predicts the presence of high-grade bladder carcinoma in patients with bladder tumors. Cent Eur J Urol. 2020;73(4):423–36. 10.5173/ceju.2020.0289. Epub 2020 Dec 3. PMID: 33552566; PMCID: PMC7848834.
25. Lin W, et al. Impact of Age at Diagnosis of Bladder Cancer on Survival: A Surveillance, Epidemiology, and End Results-Based Study 2004–2015. Cancer Control: J Moffitt Cancer Cent. 2023;30. https://doi.org/10.1177/10732748231152322.
26. Wolff RF, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51–8.

Figures

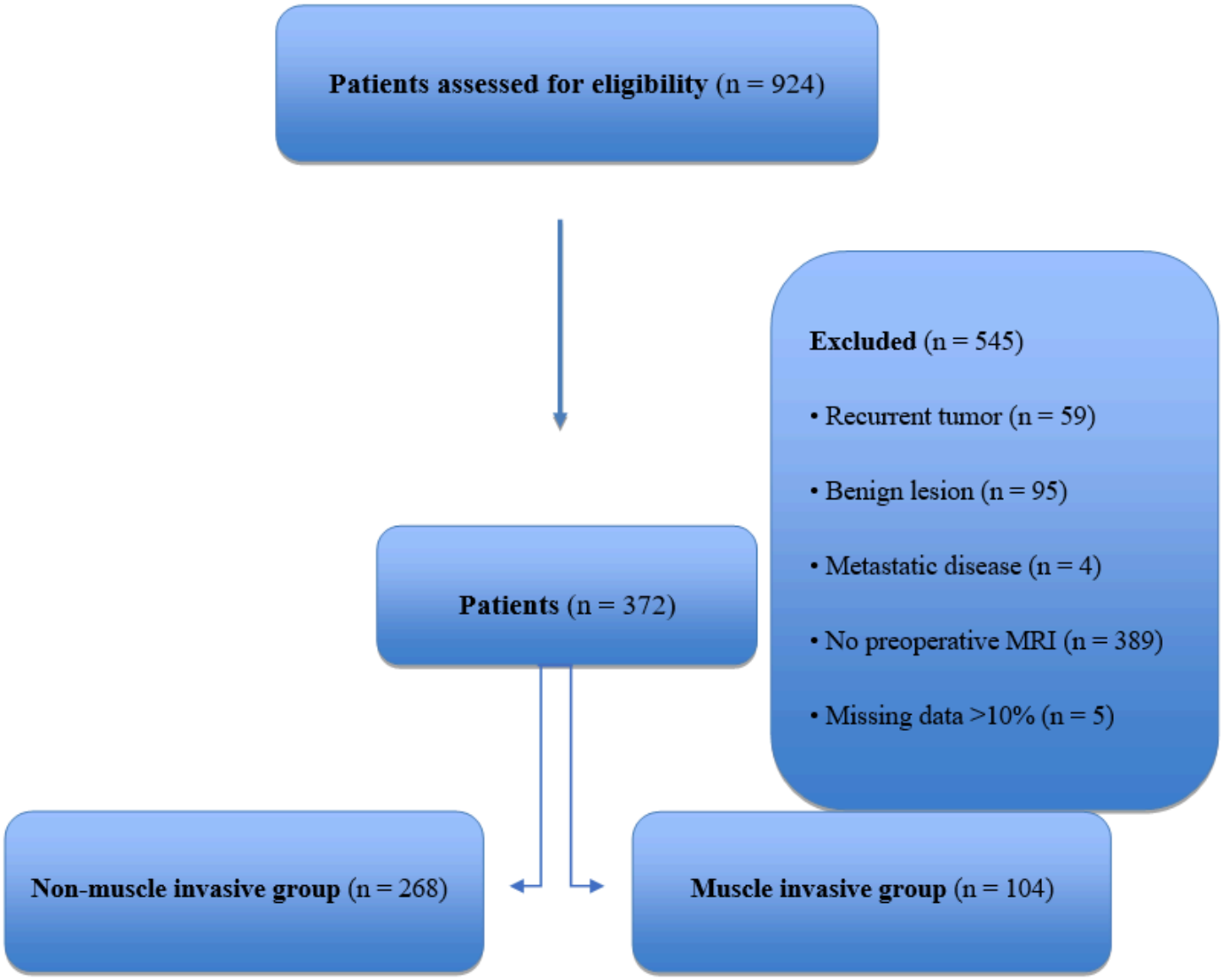


Figure 1

Flowchart of the study

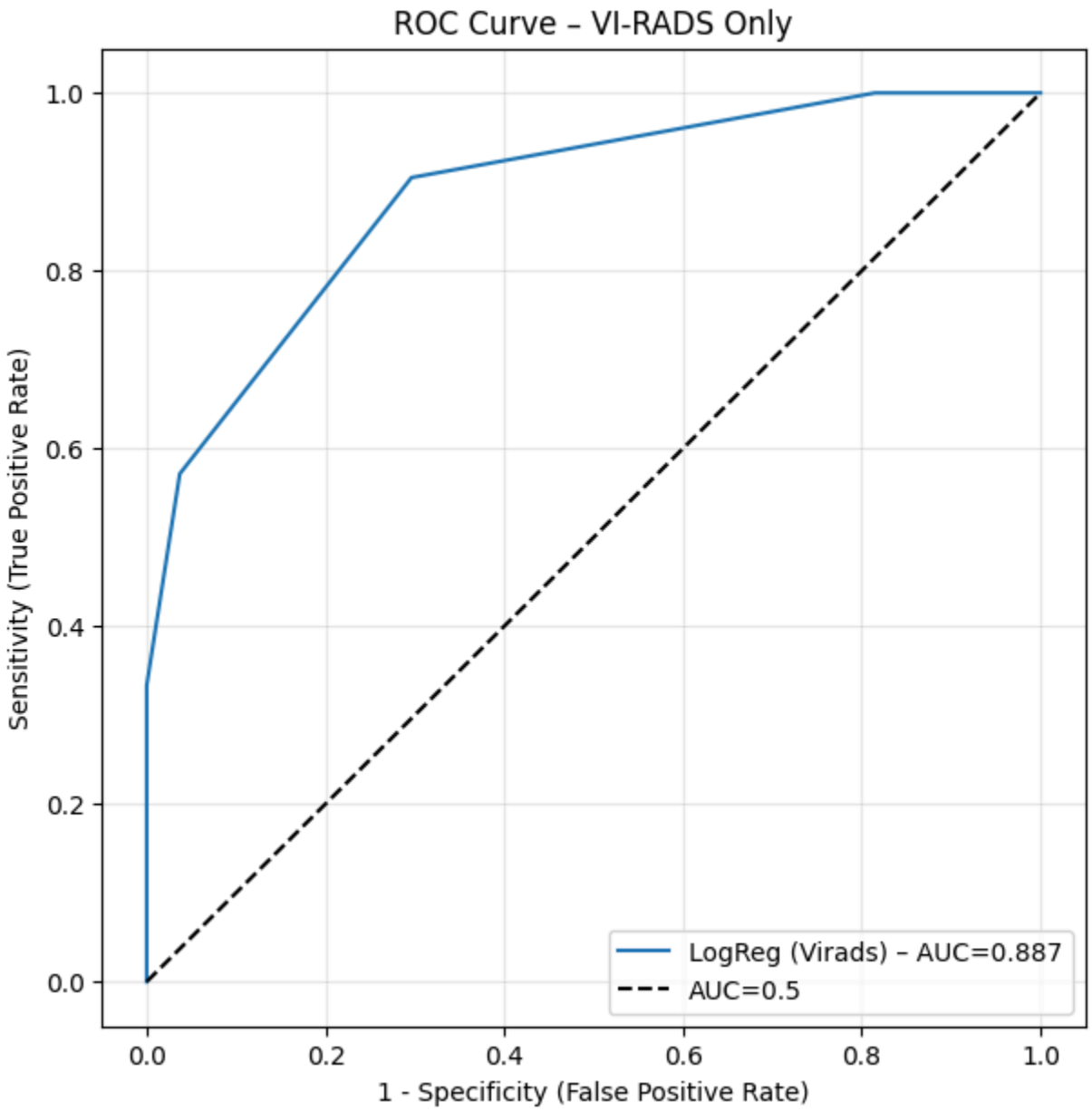


Figure 2

ROC Curve of the Model Constructed Using Only VI-RADS Score

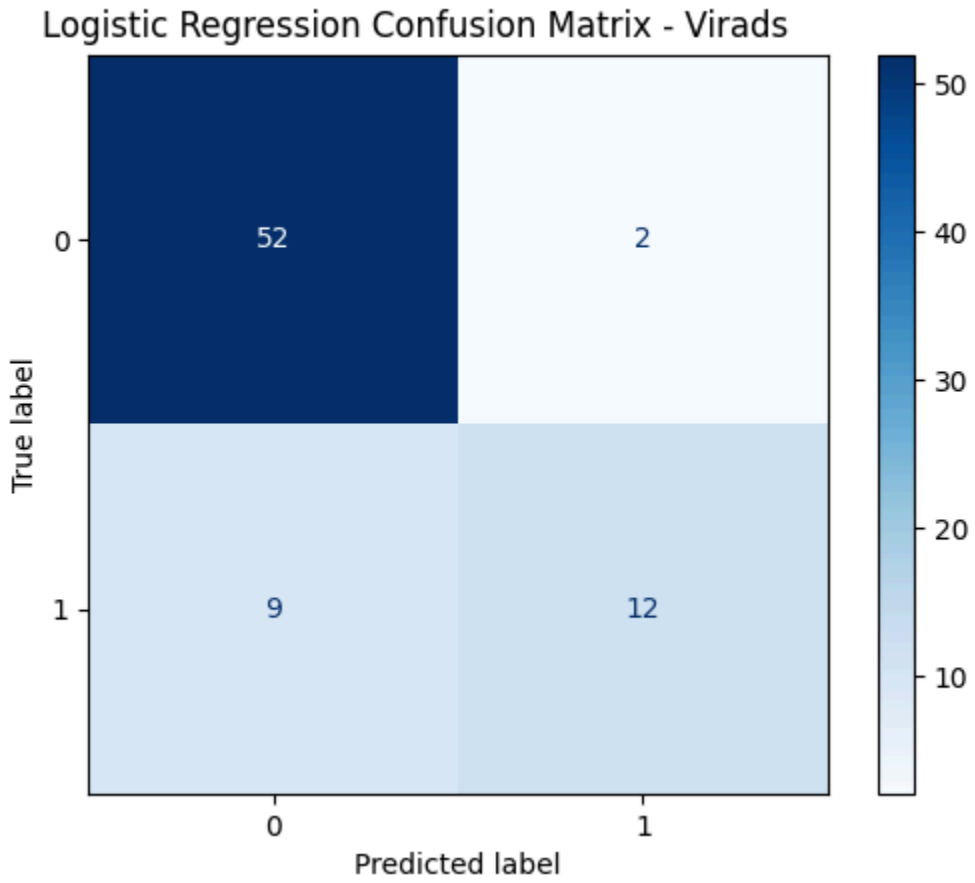


Figure 3

Confusion Matrix of the Logistic Regression Model Constructed Using Only VI-RADS Score

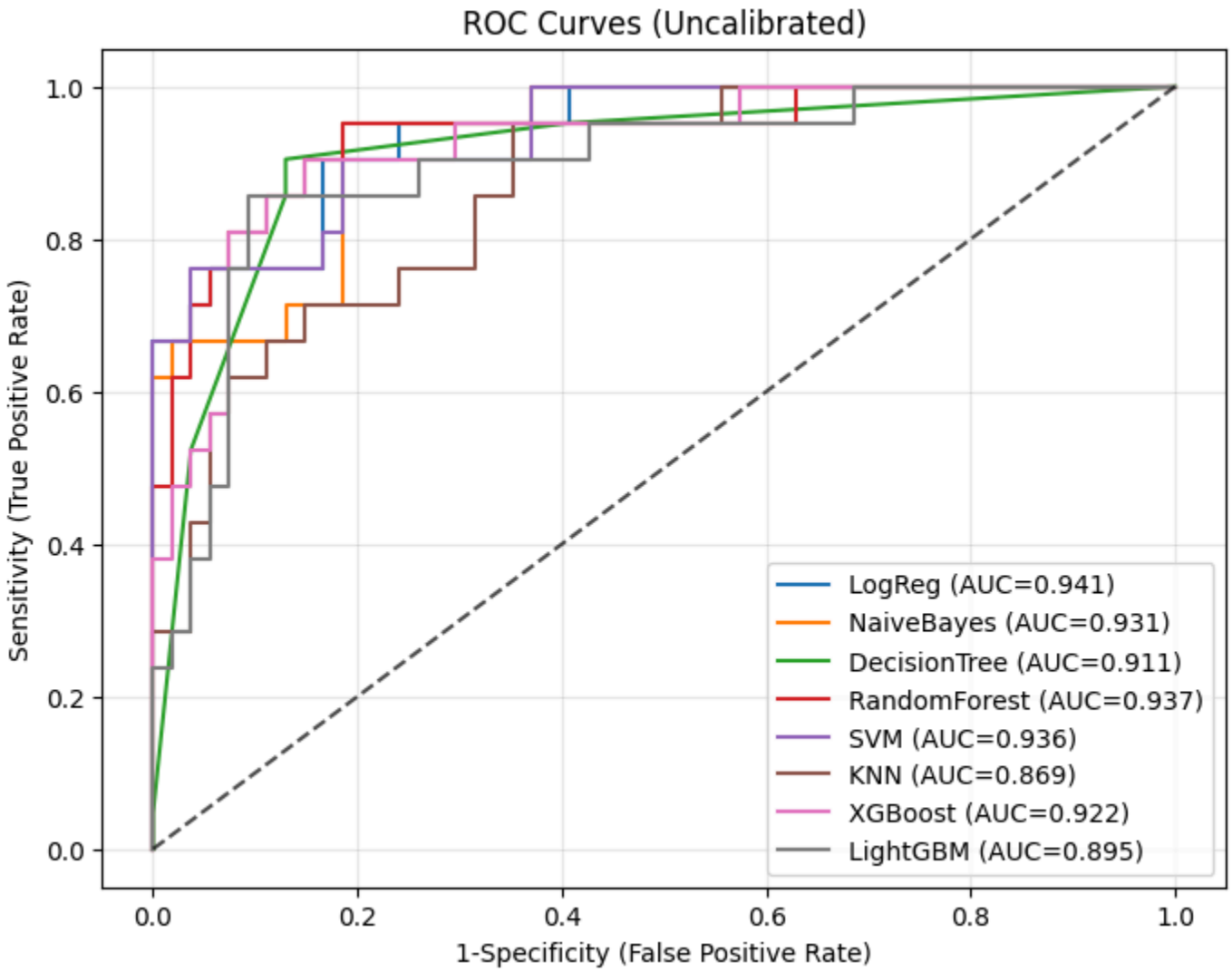


Figure 4

Uncalibrated Machine Learning Models

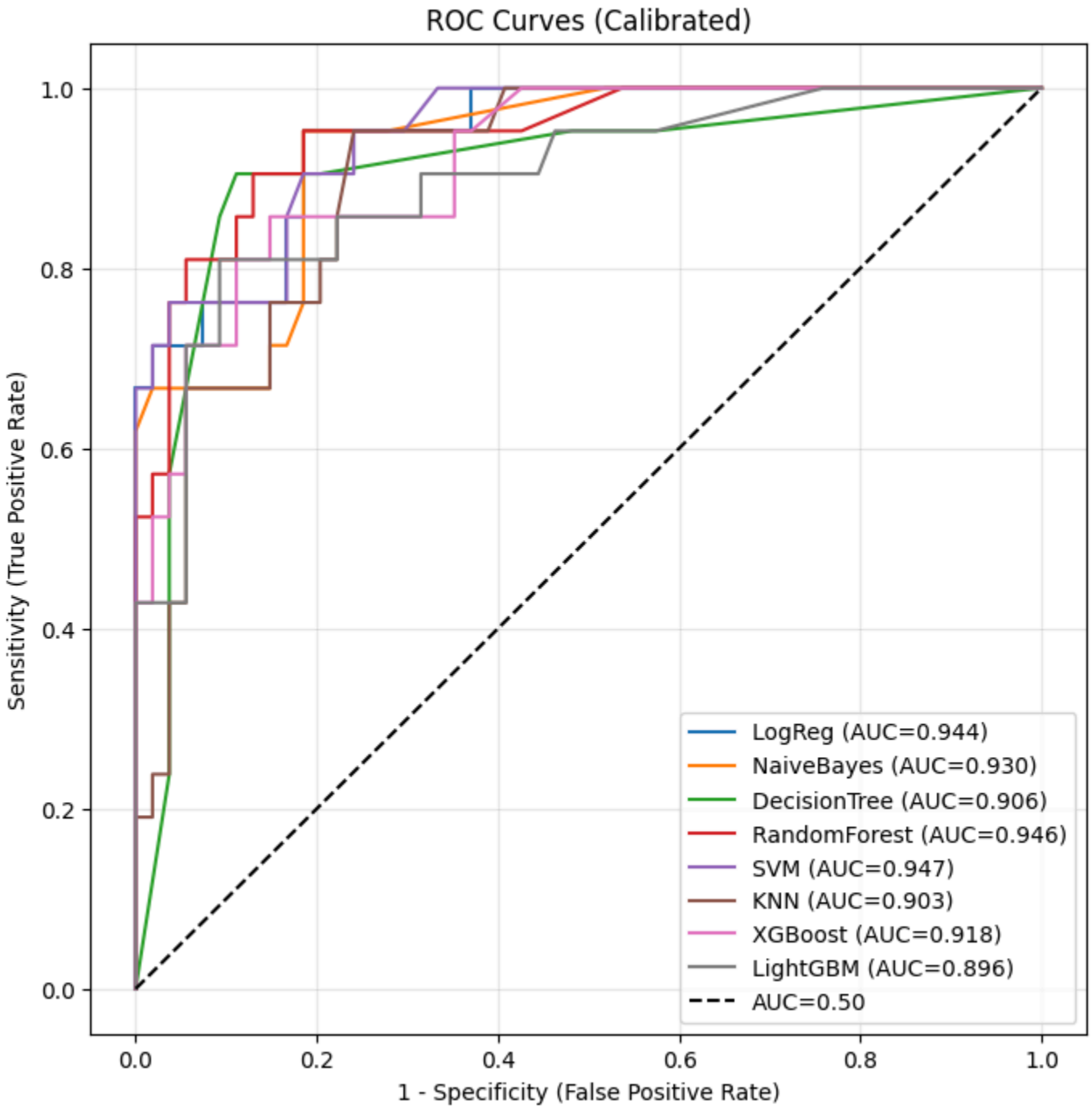


Figure 5

Machine Learning Models Calibrated Using Isotonic Regression

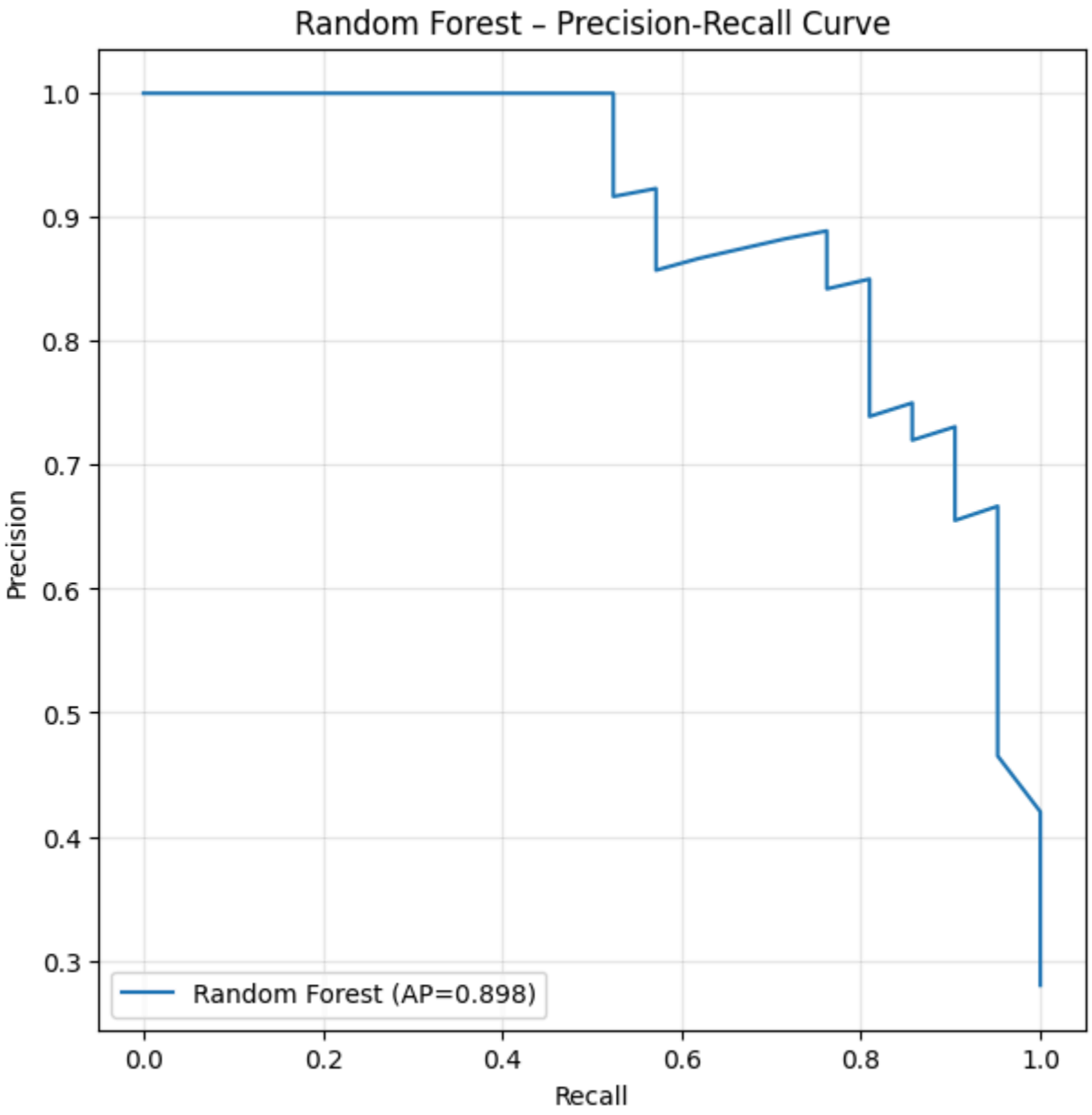


Figure 6

Precision–Recall Curve of the Random Forest Model

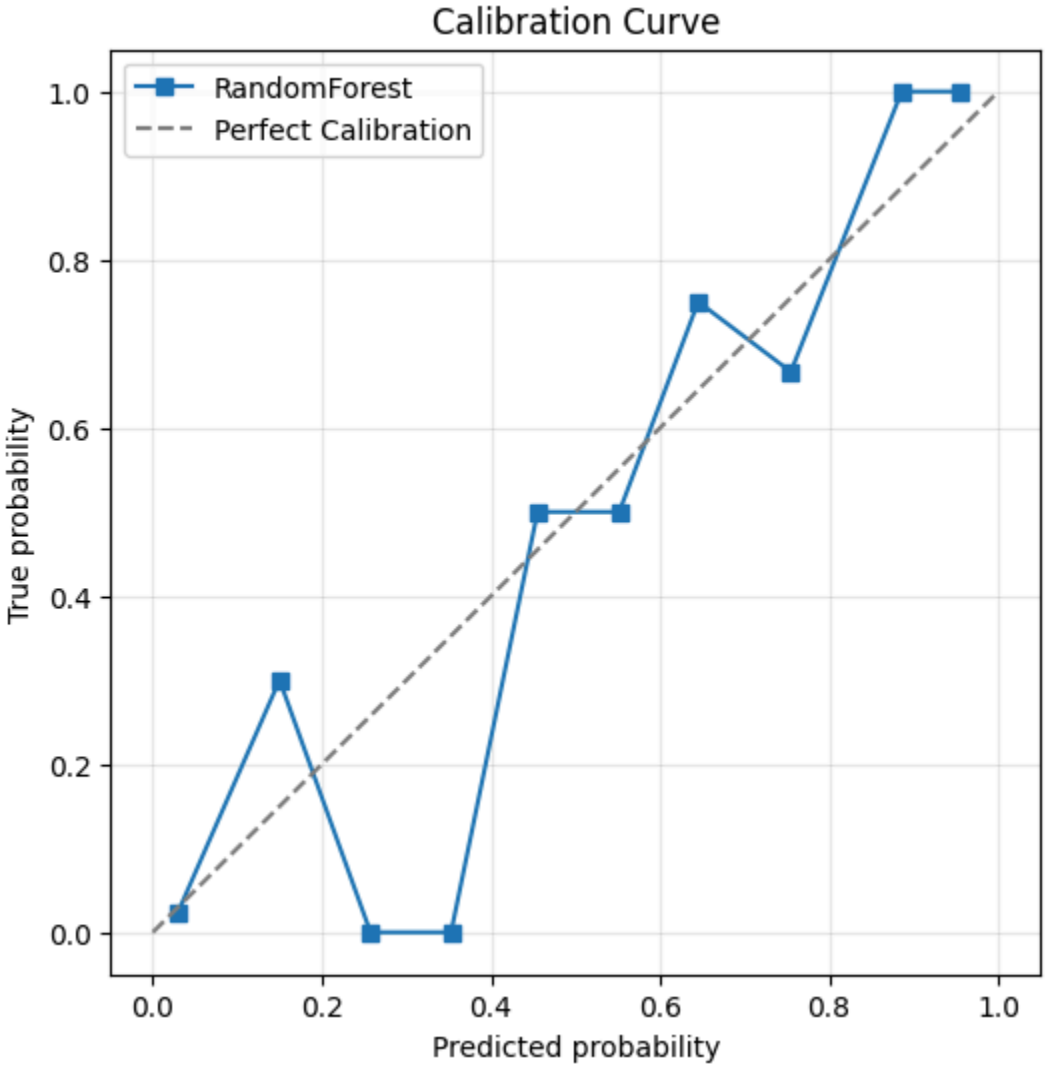


Figure 7

Random Forest Calibration Curve

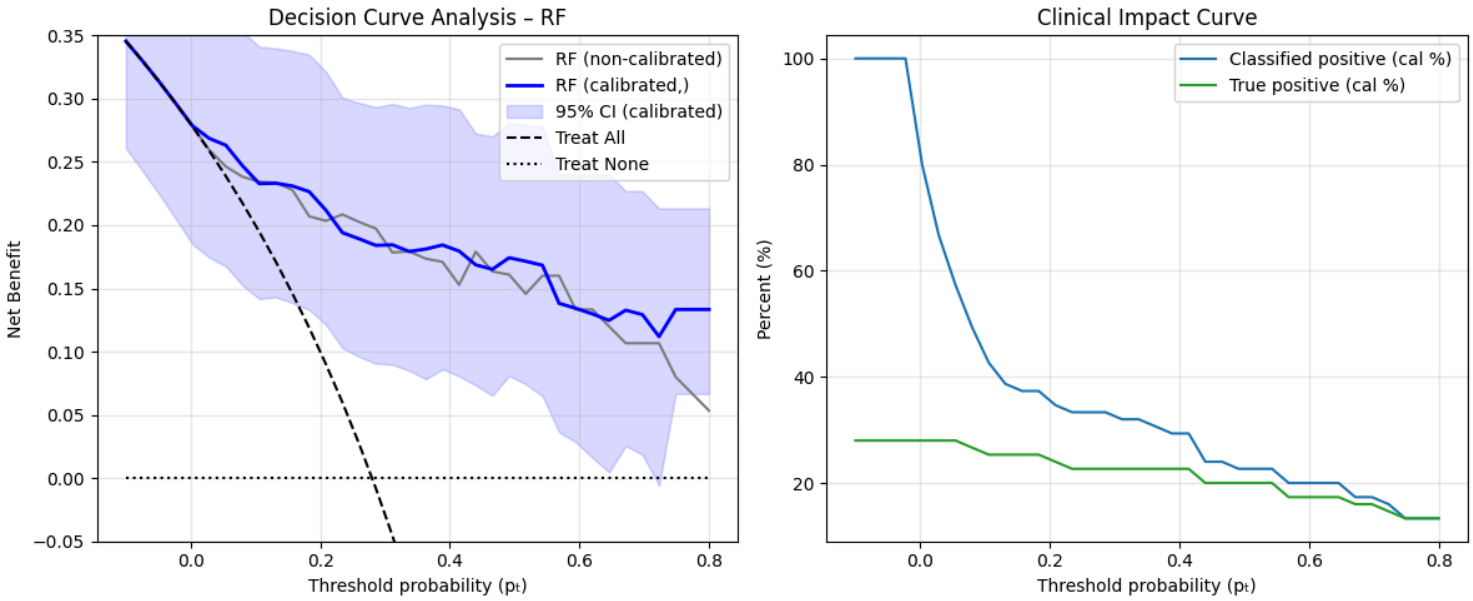


Figure 8

DCA and Clinical Impact Curves of the Random Forest Model