Machine Learning-Based Prediction of Urolithiasis Recurrence Using Patient's Clinical Data, Demography, and CT Findings

Hassan Homayoun[1], Seyed Jalaleddin Mousavirad[2], Leila Zareian Baghdadabad[1], Razman Arabzadeh Bahri[1], Iman Menbari Oskouie[1], Abdolreza Mohammadi[1], Seyed Mohammad Kazem Aghamir[1]*

Purpose: Urolithiasis is the condition of forming stones inside urinary tract with diverse shape, size, and location. The sooner urolithiasis is diagnosed, the easier it is to treat and prevent complication. This study aims to propose a method for predicting urolithiasis recurrence based on machine learning methods.

Materials and Methods: The proposed method uses clinical data, demographics, and CT findings of 4246 patients who were referred to the clinic once or multiple times within three years. The proposed method has three main phases of data engineering and pre-processing, machine learning prediction model development, and performance evaluation. In addition, the performance of six machine learning-based classifiers is evaluated by performance metric calculation, ROC curve analysis, calibration analysis, and decision curve analysis.

Results: The results of 10 independent repeats of the proposed method using a train/test split evaluation strategy reveal that the best-performing classifier is random forest with the area under the ROC curve, sensitivity, and positive predictive value of 0.64, 0.87, and 0.84, respectively. On the other hand, k-fold cross-validation: A comma is needed after "hand" and before "k-fold" evaluation strategy reveals that the best-performing classifier again is RF, with the area under the ROC curve, sensitivity, and positive predictive value of 0.63, 0.90, and 0.83, respectively. Moreover, the brier score of 0.18 shows that this classifier is well-calibrated among other evaluated classifiers.

Conclusion: This study presents a practical application of predictive machine learning methods for predicting urolithiasis recurrence with clinically acceptable accuracy compared to traditional scoring systems. To select the best classifier, six different predictive ML models have been evaluated using different performance metrics and analysis tools.

Keywords: artificial intelligence; computed tomography; machine learning; nephrolithiasis; prognosis; recurrence; urolithiasis

INTRODUCTION

Urolithiasis refers to the deposition of minerals and salts inside kidneys and formation of stones within urinary tract, including kidneys, ureters, bladder, and urethra. Thus, size, shape, and location of these stones cause diverse symptoms such as hematuria, oliguria, painful urination, dysuria, hypertension, fever, nausea, vomiting, diarrhea, loss of appetite, and abdominal/flank pain. Moreover, urolithiasis is commonly diagnosed using different imaging tests, blood tests, urine tests, and physical examinations[(1)].

Depending stone type and patient condition, the treatment of urolithiasis ranges from non-invasive to invasive options. Moreover, timely recognition of urolithiasis prevents permanent damage and complications and also increases the chance of non-expensive, non-invasive, and preventive treatment. As patients with urolithiasis are at risk and may experience stone recurrence at a rate of 50%, follow-up tests are very important. Therefore, awareness among patients with a high risk of stone recurrence improves the quality and efficiency of their therapeutic plan[(2,3)].

Urologists typically assess the likelihood of urolithiasis recurrence based on their clinical experience and established scoring systems[(4)]. However, research suggests that these scoring methods may not have a strong correlation with actual recurrence risk[(5)]. In contrast, Artificial Intelligence (AI) offers a powerful tool to enhance urologists' diagnostic and prognostic capabilities, particularly in predicting the probability of urolithiasis recurrence[(6–10)].

This study aims to propose an AI-based method for the prediction of urolithiasis recurrence based on patient's clinical data, demography, and computed tomography (CT) findings. Machine Learning (ML) models, as cutting-edge AI techniques are able to integrate complex high-dimensional data, find non-linear relationships between diverse variables, and derivation of final decision [(11)]. Hence, the proposed method utilizes ML models among the most powerful predictive techniques to predict urolithiasis recurrence.

MATERIALS AND METHODS

Patients

Demography, patient's clinical data (including symptoms and interventions), and abdominopelvic CT findings of 4246 patients who referred once or multiple

[1]Urology Research Center, Tehran University of Medical Sciences, Tehran, Iran.
[2]Department of Computer and Electrical Engineering, Mid Sweden University, Sweden.
*Correspondence: Urology Research Center, Sina Hospital, Hassan Abad Sq., Imam Khomeini Ave., Tehran, Iran.
Tel: (+9821) 6634 8560. Fax: (+9821) 6634 8561. Email: mkaghamir@tums.ac.ir.
Received January 2025 & Accepted November 2025

Table 1. The selected features with their corresponding valid values.

Feature No.	Feature Name	Valid Values	Description	Category
1	Sex	• Male • Female	Patient's sex	Demography
2	Age	[Valid integer]	Patient's age in years	Demography
3	Height	[Valid integer]	Patient's height in centimeters	Demography
4	Weight	[Valid integer]	Patient's weight in kilograms	Demography
5	Economic-Status	• High • Upper Medium • Medium • Lower Medium • Low	Economic status of family	Demography
6	Housing-Status	• Personal Housing • Rental Housing • Government Housing	Housing status of family	Demography
7	Clinical-Symptoms	• Vulvodynia • Hematuria • Oliguria • Painful Urination • Dysuria • Hypertension • Fever • Nausea • Vomiting • Diarrhea • Loss of Appetite • Weight Loss • Abdominal and Flank Pain • Testicle Pain	Patient's symptoms	Patient's Clinical Data
8	Season	• Spring • Summer • Autumn • Winter	Season in which the patient experienced symptoms	Demography
9	Referring-Reason	• Clinical Symptoms • Blood Testing • Urine Cultivation • Accidentally	The reason in which the patient refers for treatment	Patient's Clinical Data
10	Stone-Count-RK	[Valid integer]	Stone count in the right kidney in centimeters	CT Findings
11	Stone-Count-LK	[Valid integer]	Stone count in the left kidney in centimeters	CT Findings
12	Observing-Hydronephrosis	• Yes • No	Status of hydronephrosis occurrence	CT Findings
13	Involved-Kidneys	• Right kidney • Left kidney	Kidneys in which the stone is observed	CT Findings
14	Stone-Place	• Upper Calyx • Anterior Upper Calyx • Posterior Upper Calyx • Medial Upper Calyx • Lateral Upper Calyx • Middle Calyx • Anterior Middle Calyx • Posterior Middle Calyx • Medial Middle Calyx • Lateral Middle Calyx • Inferior Calyx • Anterior Inferior Calyx • Posterior Inferior Calyx • Medial Inferior Calyx • Lateral Inferior Calyx • Ureter • Proximal Ureter • Middle Ureter • Distal Ureter • Anterior Ureteral Valve • Posterior Ureteral Valve • Renal Pelvis • Bladder	The places in which stones are observed	CT Findings
15	Largest-Stone-Size	• < 2 • [2 - 5] • [6 - 10] • [11 - 15] • [15 - 20] • [20 - 30] • [30 - 50] • > 50	Largest stone size in millimeters	CT Findings
16	Follow-up-Treatment-Intervention-Methods	• Observation • Drug Intervention • Surgical Intervention	The used intervention methods in the follow-up	Patient's Clinical Data

Abbreviations: RK, right kidney; LK, left kidney.

times to the urology clinic of Sina hospital (Tehran, Iran) during the last three years are fetched from the national database of urolithiasis registry with the approval by Tehran University of Medical Sciences institutional review board under ethic code (IR.TUMS.MEDICINE.REC.1403.231). Among these patients, 3417 experienced recurrences within two years after observing stone-free conditions with the confirmation of an expert

Table 2. Hyperparameters of six classifiers used in the proposed method, including NB, SVC, NN, RF, GB, and KNN

Classifier	Hyper-parameters
RF	splitting criterion = 'gini'; maximum depth of forest = 11; method for selecting maximum number of features = 'sqrt'; number of base learners = 500; minimum samples split = 3; minimum samples leaf = 1;
GB	learning rate = 0.01; splitting criterion = 'squared_error'; maximum depth = 15; number of base learners = 400; method for selecting maximum number of features = 'log2'; minimum samples split = 5;
NN	activation function = 'relu'; size of first, second, and third hidden layers = (60, 30, 10); solver = 'adam'; early stopping = active; number of iterations with no change to active early stopping = 10;
SVC	C = 100; gamma = 1; kernel = 'rbf';
NB	no parameter to tune;
KNN	number of neighbors = 17; distance metric = 'minkowski'; metric parameter = 2;

urologist. Moreover, 829 patients have not experienced recurrence and remain in stone stone-free condition for more than two years. In this study, CT scans are the most important test for assessing recurrence. Due to the nationwide referral nature of urology clinic of Sina hospital for urolithiasis patients, the dataset is imbalanced with a strong bias towards patients experiencing recurrence.

To ensure the most relevant information is considered for prediction, sixteen features are selected for each patient based on relevancy and availability. More details on these features can be found in **Table 1**.

Prediction Method

To predict the risk of urolithiasis recurrence based on patient's clinical data, demography, and CT findings, an ML-based method is proposed. The core of ML-based prediction methods is ML models, which can learn and identify patterns related to a specific task from the training examples (data) and perform that task on unseen data without explicit programming. Supervised learning, an ML paradigm where models learn from labeled data to uncover patterns and map inputs to desired outputs, forms the core of the proposed method for predicting urolithiasis recurrence. As it is depicted in **Figure 1**, the input of the proposed supervised ML model is the data of patients who experienced urolithiasis, and the output is whether the patient will experience recurrence or not, in the future. The proposed method of predicting urolithiasis recurrence has three main phases of data engineering and pre-processing, prediction model construction, and performance evaluation, which are explained in the following sections.

Data Engineering and Pre-processing

As we are using a real-world dataset in this study, it is usual to come up with data quality issues such as missing values, outliers, and scale inhomogeneity. In order to address these issues and make the data ready for prediction, some data engineering and pre-processing tasks are performed. Hence, a quick exploratory data analysis (EDA) is conducted to get insights into the data and decide on the required pre-processing and data engineering tasks. This section explains performed pre-processing tasks.

• Missing Value Imputation

In the process of entering relevant data into the national database of urolithiasis registry, sometimes, some features are missed for any reason. In order to fill missed values with an appropriate value, the nearest neighbor imputation is employed. This imputation method approximates missing values by considering the neighboring data points. In this study, three neighbors of each data point with a missing value are considered[12]. The reason for this choice is that considering three neighbors produces the minimum Bayes error, which is the minimum possible error, theoretically[13]. Worth mentioning that the rate of missing values for features that go under the imputation process does not exceed 10%.

• Outlier Correction

In the process of EDA, it has been observed that some features have extreme values. These extremes prevent proper learning of the statistics of these features by ML models. Therefore, to correct these out-of-bound values, minimum extremes and maximum extremes are replaced with 1th and 99th percentiles of each feature, respectively.

• Feature Scaling

In many ML models, in the process of learning model parameters, feature values are either multiplied by a value that somehow determines their importance or the distance between samples in the feature space is measured. In order to prevent ML models from becoming biased toward features with higher ranges in the training process, all numerical values are scaled into the range of [0, 1] by using min-max normalization. Equation 1 describes how a new scaled value for a typical feature is obtained.

$$value_{new} = \frac{value_{old} - min(feature)}{max(feature) - min(feature)} \qquad (1)$$

• One-hot-encoding

As many of the features in the dataset are categorical, in order not to ML models treat them as numerical features, they are converted to a categorical encoding known as one-hot-encoding. As illustrated in **Figure 2**,

Table3. Average performance evaluation of all classifiers for 10 repeats of the train/test split strategy in terms of ACC, SEN, SPE, PPV, NPV, AUC ROC, and BS.

Classifier	ACC (std.)	SEN (std.)	SPE (std.)	PPV (std.)	NPV (std.)	AUC-ROC (std.)	BS (std.)
RF	0.76 (0.02)	0.87 (0.02)	0.30 (0.02)	0.84 (0.01)	0.35 (0.04)	0.64 (0.02)	0.19 (0.01)
SVC	0.72 (0.01)	0.86 (0.01)	0.13 (0.01)	0.80 (0.00)	0.19 (0.02)	0.51 (0.02)	0.25 (0.01)
KNN	0.63 (0.02)	0.69 (0.02)	0.41 (0.04)	0.83 (0.01)	0.24 (0.02)	0.57 (0.02)	0.25 (0.01)
NN	0.68 (0.02)	0.76 (0.03)	0.31 (0.05)	0.82 (0.01)	0.24 (0.02)	0.55 (0.03)	0.25 (0.02)
GB	0.78 (0.01)	0.93 (0.01)	0.15 (0.03)	0.82 (0.00)	0.35 (0.04)	0.62 (0.02)	0.20 (0.01)
NB	0.63 (0.02)	0.67 (0.02)	0.49 (0.03)	0.84 (0.01)	0.26 (0.02)	0.59 (0.02)	0.23 (0.00)

Table4. Performance evaluation of all classifiers using 10 fold cross validation in terms of SEN, SPE, PPV, NPV, AUC ROC, and BS

Classifier	SEN (95% CI)	SPE (95% CI)	PPV (95% CI)	NPV (95% CI)	AUC-ROC (95% CI)	BS (95% CI)
RF	0.90 (0.89 -0.91)	0.22 (0.21-0.23)	0.83 (0.81-0.84)	0.34 (0.30-0.38)	0.63 (0.60-0.65)	0.18 (0.18-0.19)
SVC	0.91 (0.90-0.92)	0.08 (0.07-0.09)	0.80 (0.79-0.82)	0.18 (0.14-0.21)	0.53 (0.50-0.55)	0.21 (0.20-0.22)
KNN	0.47 (0.45-0.49)	0.65 (0.63-0.66)	0.85 (0.83-0.86)	0.23 (0.21-0.25)	0.57 (0.55-0.59)	0.30 (0.29-0.30)
MLP	0.81 (0.79-0.82)	0.31 (0.29-0.32)	0.83 (0.81-0.84)	0.28 (0.25-0.31)	0.57 (0.55-0.60)	0.23 (0.22-0.24)
GB	0.95 (0.95-0.96)	0.12 (0.11-0.13)	0.82 (0.81-0.83)	0.39 (0.33-0.45)	0.61 (0.59-0.63)	0.16 (0.15-0.17)
NB	0.13 (0.12-0.15)	0.86 (0.85-0.88)	0.80 (0.77-0.84)	0.19 (0.18-0.21)	0.54 (0.51-0.56)	0.71 (0.69-0.72)

Abbreviations: CI, Confidence Interval.

a categorical feature with I valid values is converted to I binary features.

• Up-sampling

As we have described in the patient section, the dataset is highly imbalanced toward patients who experienced recurrence. In order to prevent ML models from being biased toward the major class of patients, the minor class is up-sampled by employing the adaptive synthetic sampling (ADASYN) up-sampling method[14]. This method generates synthetic samples for the minor class by considering the distribution of the minor class and focusing on sparse areas of feature space. Worth mentioning that up-sampling is applied only to the training data during the training phase and is never used for test data in the evaluation phase.

Prediction Model Construction

After the data engineering and pre-processing phase, the data are ready to be fed to the ML training algorithm for learning model parameters. This study explores the effectiveness of six different classification models, including Naive Bayes (NB), Support Vector Classifier (SVC), Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbor (KNN), and Neural Network (NN) for making predictions of urolithiasis recurrence. NB classifier attempts to model the distribution of the data by holding the assumption of independence between features[15]. SVC tries to find a decision boundary with maximum margin for better separation of patients into two groups[16]. RF decides by employing multiple base decision trees, which are trained based on different subsets of training data[17]. GB also builds an ensemble of weak classifiers by sequential inclusion of weak learners for overall performance improvement[18]. KNN classifier, as a lazy learning method, predicts by looking at its k nearest neighbors[19]. NNs, which mimic brain structure, attempt to perform non-linear mapping from input to output in a layer-by-layer fashion, and finally make a classification decision[20]. In order to predict the risk of urolithiasis recurrence for new patients, three tasks of hyperparameter tuning, training, and calibration must be completed for each classifier. In the following sub-sections, the tasks of hyper-parameter tuning, training, and calibration of these six classifiers are described.

• Hyperparameter Tuning:

The purpose of hyperparameter tuning is to determine the configuration of classifiers for better encoding inductive biases related to the problem at hand, highly exploiting training data, and ultimately producing a quality input-output mapping. For all of the classifiers, the best hyperparameters are found using a grid search procedure [21]. Grid search explores different combinations of valid values of hyperparameters and selects the best set suitable for the task at hand. **Table 2** summarizes the hyperparameters of all the above-mentioned classifiers.

• Training:

After tuning hyperparameters of each classifier, they are trained using 80% of the data. The goal of training a classifier is to learn and approximate the underlying distribution of the data using different algorithms relevant to each specific classifier. As mentioned previously, before the training process, as the dataset is highly imbalanced, the minor class is up-sampled so as not to bias the training models toward the major class.

• Calibration:

In order to force classification models to provide more reliable and confident probabilities for their predictions, a sigmoid scaling on the output of prediction models is

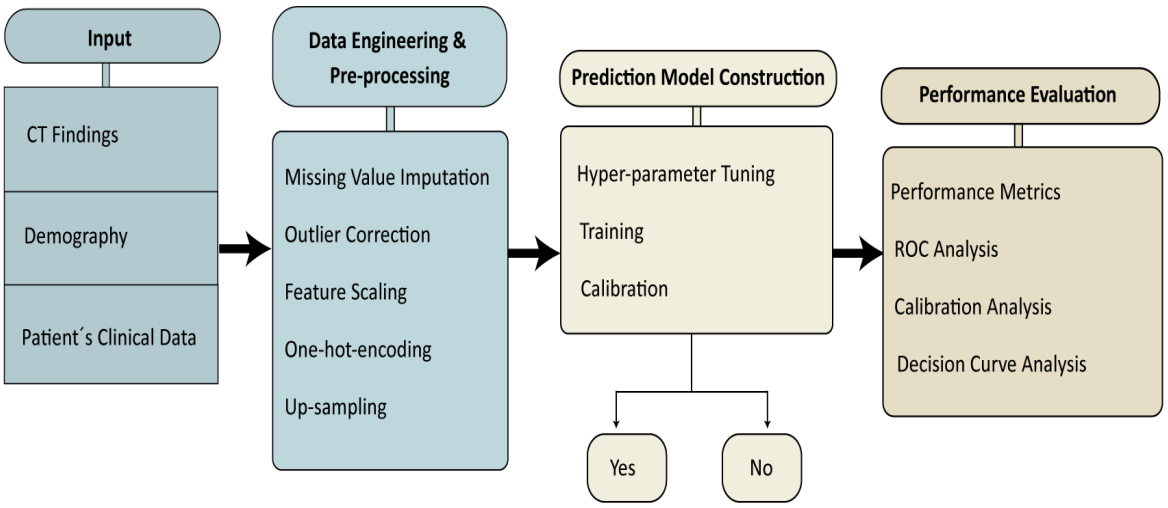

Figure 1. Block diagram of the proposed method for predicting urolithiasis recurrence, including main phases and corresponding tasks.

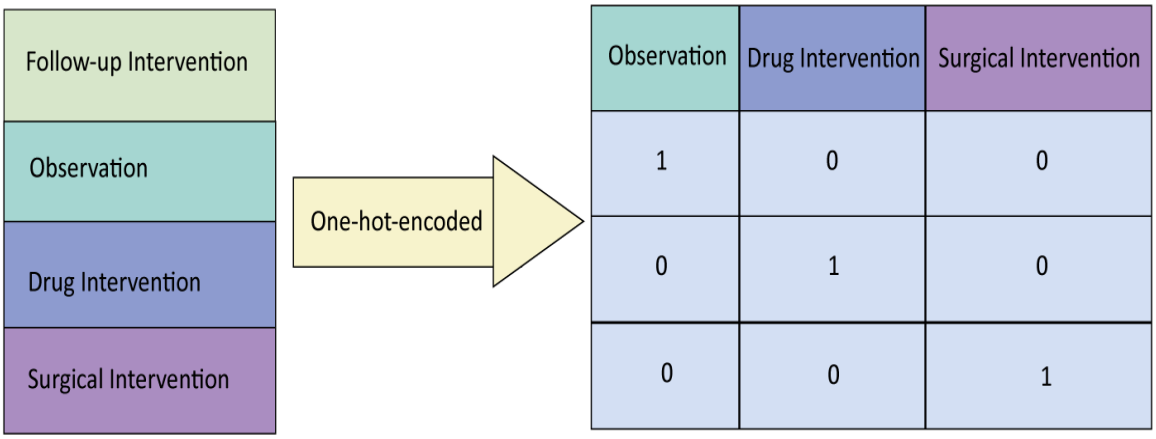

Figure 2. One-hot-encoding: Follow up Intervention feature with three valid values converted into three binary features (Observation, Drug Intervention, Surgical Intervention).

applied to map scores of predictions into more reliable probabilities [22].

Performance Evaluation of the Prediction Models

To evaluate the performance of the proposed method, two strategies of train/test splitting and k-fold cross-validation are employed. In the train/test splitting strategy, all six classifiers are trained on 80% of the data and tested on the remaining 20%. This 20% of the data, as a test set, has never been seen by the model in the training phase of model construction and is only used to calculate performance measures in the test/evaluation phase. Before the calculation of these performance measures, the results of classification models are summarized in a table-like structure called confusion matrix. As it is depicted in **Figure 3**, a confusion matrix breaks down classification results into four categories of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP and TN are samples that the prediction model correctly identified as positive and negative, respectively. FP and FN are samples that the prediction model wrongly identified as positive and negative, respectively. In addition to the train/test splitting strategy, a 10-fold cross-validation strategy is employed to assess the performance of the proposed method. In this strategy, 10% of the data as a training fold is set aside for testing, and the remaining data is used for training. This process, which is repeated 10 times, allows us to use all the data for both training and testing. To assess how well the prediction models performed, several performance metrics such as Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Positive Predictive Value (PPV), and Negative Predictive Value (NPV) are calculated as Equations 2 to 7 explained. ACC measures the overall performance of models by measuring the correct prediction of the model. SEN and

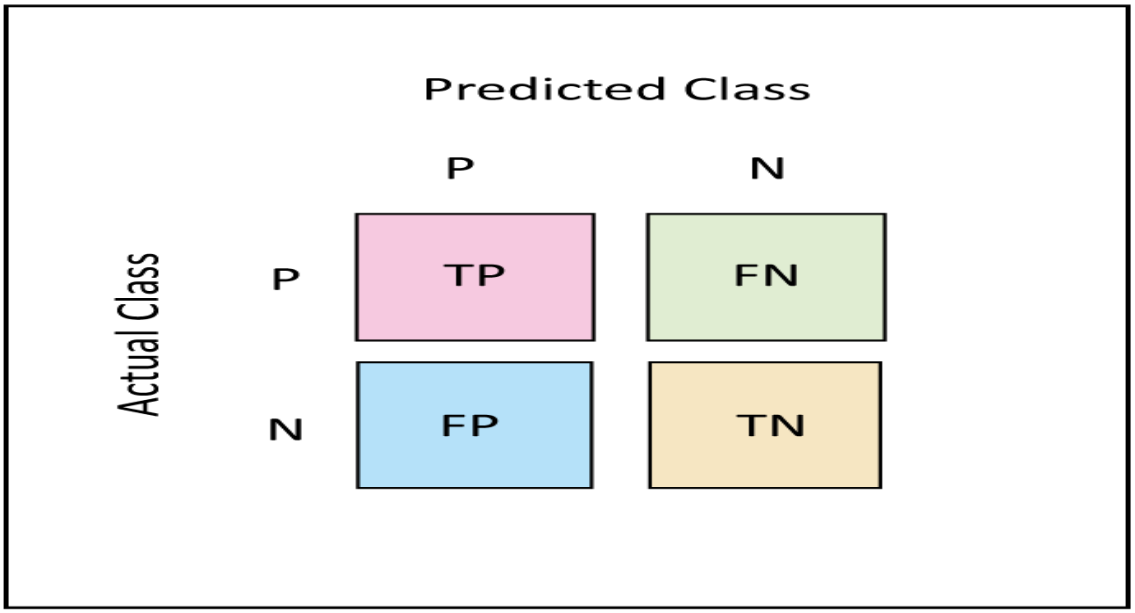

Figure 3. Confusion matrix illustrating TP, TN, FP, and FN.

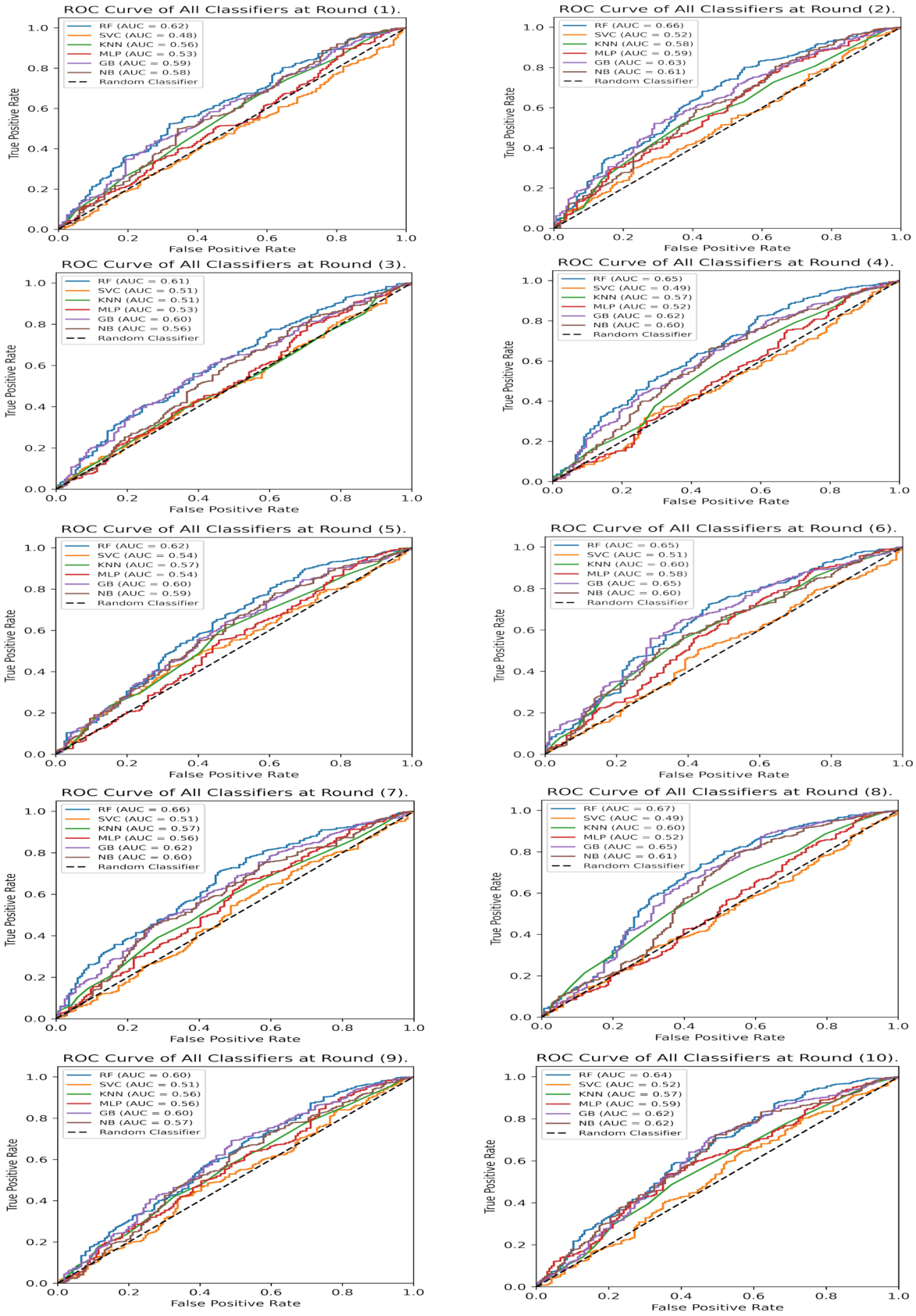

Figure 4. ROC curves of all classifiers across 10 independent repeats.

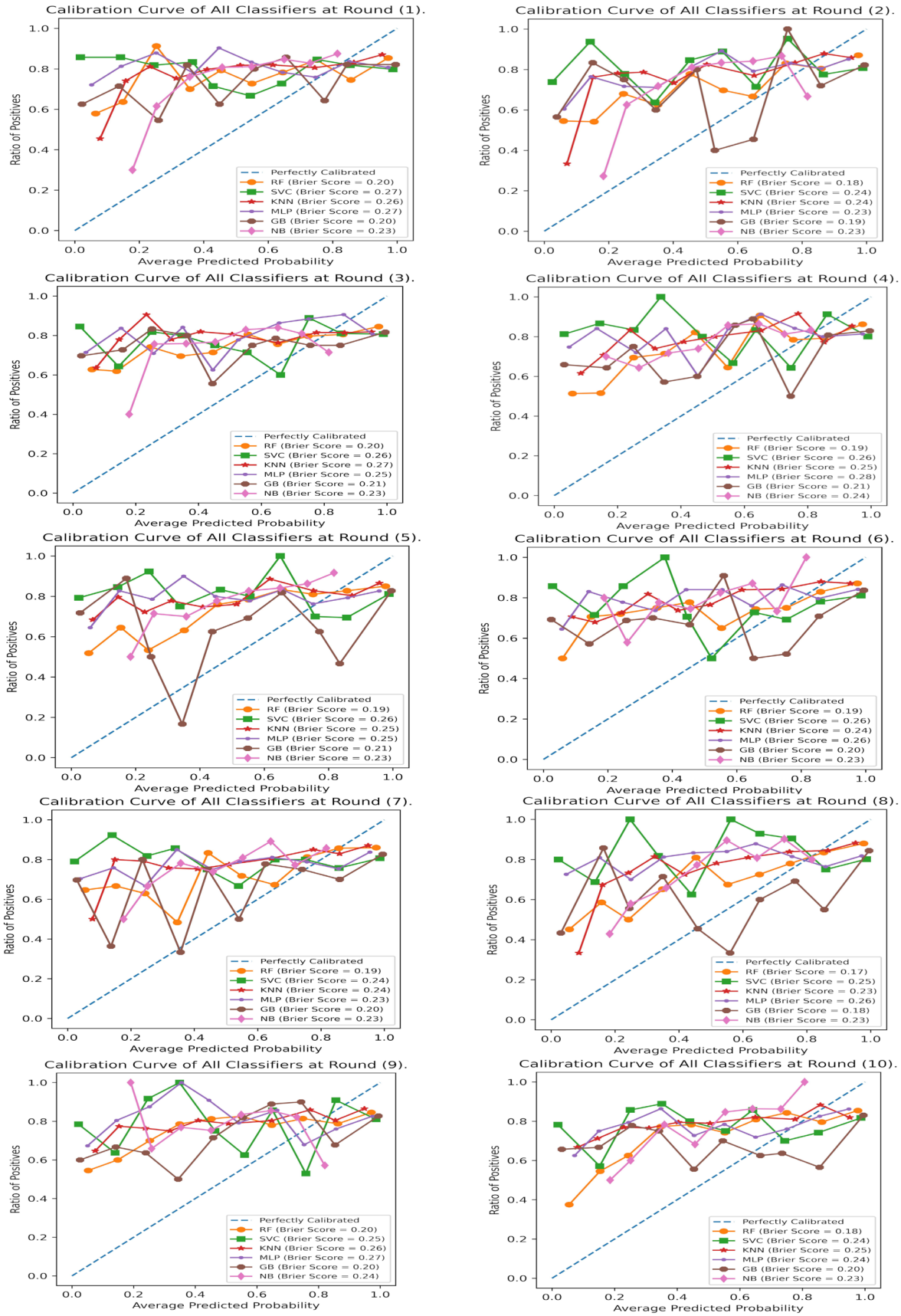

Figure 5. Calibration curves of all classifiers across 10 independent repeats of the train/test split evaluation.

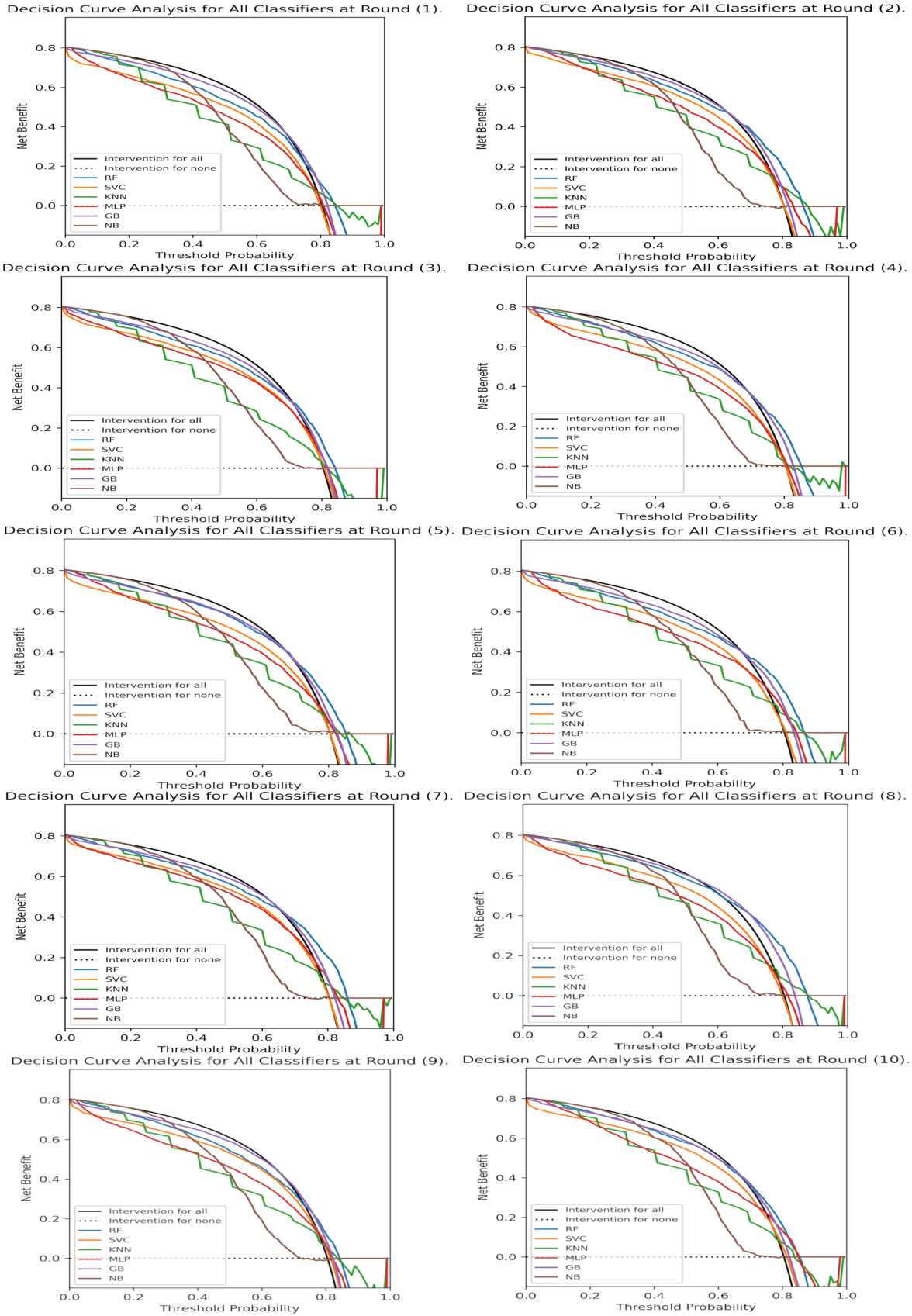

Figure 6. Decision curves of all classifiers across 10 independent repeats of the train/test split evaluation.

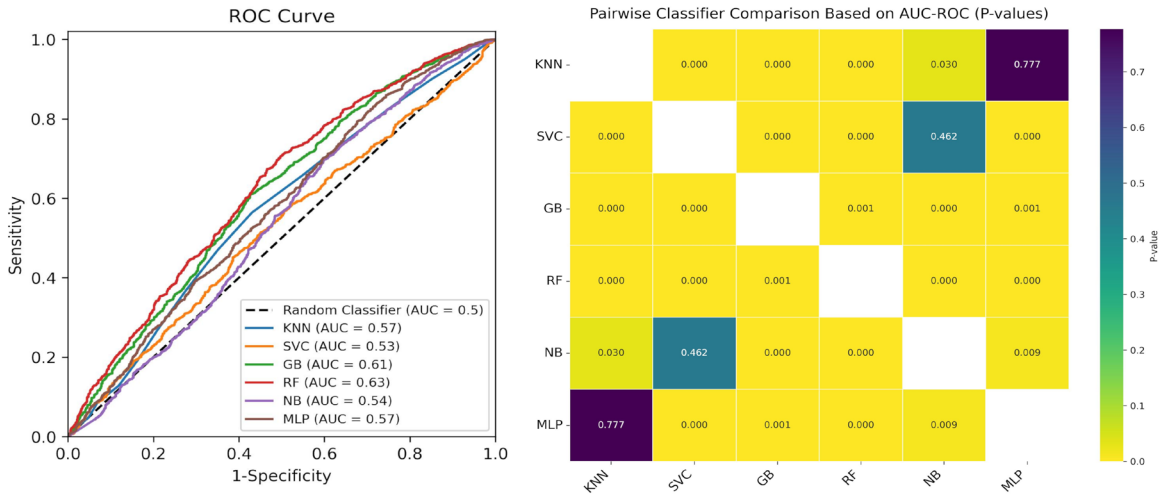

Figure 7. Left: ROC plot under 10-fold cross validation. Right: Statistical comparison of classifiers by AUC ROC.

SPE measure the model's ability to correctly identify positive cases and negative cases, respectively. PPV measures the exactness of the model by calculating the proportion of correct positive predictions, while NPV focuses on the proportion of negative predictions that are correct. The valid value of the mentioned performance measures ranges from zero (worst model) to one (best model).

In addition to the mentioned performance metrics, Receiver Operation Characteristics (ROC) analysis is performed to have a more robust and informative method of evaluation of the prediction models[23]. ROC curve actually is a graphical tool that plots a classifier's True Positive Rate (TPR) (Sensitivity) on the y-axis against the False Positive Rate (FPR) (1-Specificity) on the x-axis. As it is explained by Equations 3 and 7, TPR represents the proportion of positive cases the model correctly identified, while FPR represents the proportion of negative cases incorrectly classified as positive. Measuring the Area Under the ROC Curve (AUC-ROC) estimates the overall performance of the proposed prediction method with focusing on increasing TPR and decreasing FPR. This area, which ranges from zero (worst model) to one (best model), indicates whether the model performs better than a random classifier (AUC-ROC = 0.5) or not. Indeed, AUC-ROC allows to have a fair comparison of different classifiers.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$SEN, TPR = \frac{TP}{TP+FN} \quad (3)$$

$$SPE = \frac{TN}{TN+FP} \quad (4)$$

$$PPV = \frac{TP}{TP+FP} \quad (5)$$

$$NPV = \frac{TN}{TN+FN} \quad (6)$$

$$FPR = \frac{FP}{FP+TN} \quad (7)$$

With the aim of evaluating confidence of classifiers (how well the probabilistic output of classifiers is compared to the actual output), calibration curve analysis is performed[24]. For each classifier, the predicted probability in the x-axis against the observed proportion of positive outcomes for specific predicted probability bins in the y-axis is plotted. It is expected that the curve of a well-calibrated classifier lies in a diagonal line in 45 angle degrees of the plot. Moreover, each plot is summarized via Brier Score (BS), which measures the mean square difference between predicted probabilities and actual outcome as described by equation 8. In this equation, N, f^i, and o^i, are total number of predictions, predicted probability for the i^{th} sample, and actual outcome of the i^{th} sample, respectively.

$$BS = \frac{1}{N}\sum_i^N (f_i - o_i)^2 \quad (8)$$

In order to understand the real-world impact of the proposed method, by considering the potential benefits and harms of taking actions based on the classifier's prediction, a Decision Curve Analysis (DCA) is performed [25]. DCA allows us to plot the net-benefit of the prediction model against the minimum predicted probability of the event at which we decide to take action. As it is described by Equation 9, net-benefit (NB), as a combined metric, considers both the benefit of correctly classifying true positives and the harm of incorrectly classifying false positives. In this equation, n and t are the total population and threshold, respectively. In DCA analysis, net-benefits of default strategies of intervention for all and intervention for non (zero net-benefit) are plotted, which allows better evaluation.

$$NB = \frac{TP}{n} - \frac{FP}{n}.\left(\frac{t}{1-t}\right) \quad (9)$$

RESULTS

After implementing the proposed method, we trained all prediction models based on the training data. To assess the performance of prediction models, all of them

are evaluated on the test data, and all mentioned performance metrics are calculated. The evaluation process in train/test splitting strategy was repeated 10 times independently to be more reliable. For each round of evaluation, both training and test data are partitioned randomly. Finally, the results are averaged for 10 repeats and summarized in Table 3. In terms of recognition power of both positive and negative cases, the best classifier is GB with an accuracy score of 0.78 (0.01). The GB classifier is also the most sensitive classifier with a sensitivity score of 0.93 (0.01). The results demonstrated that the most specific classifier is NB with a specificity score of 0.49 (0.03). The results also demonstrated that the most exact classifiers are RF and NB, with a PPV of 0.84 (0.04). NPV of 0.35 (0.04) confirmed superiority of RF and GB over other classifiers in terms of reliability of negative prediction.

In order to measure the performance of the proposed method by considering the trade-off of catching true positive cases and avoiding false positives, a ROC curve analysis is performed. As it is depicted in **Figure 4**, for all 10 rounds of experimentation, RF stays on top of other classifiers. Moreover, as it is shown in **Table 3**, the average ROC-AUC of 0.64 (0.02) confirms this superiority. The results also show second-best performing classifier in terms of ROC-AUC is GB. As both RF and GB classifiers are from the same family, this reveals the ability of tree-based classifiers for better modeling of these types of data. In addition, the worst classifier is SVC, which is close to the random classifier.

In addition to the analysis of ROC curves, calibration curve analysis and brier score calculation for all classifiers in 10 independent repeats are performed. As illustrated in Figure 5, because calibration curves are above the diagonal line of a well-calibrated classification model, all models are almost overconfident, which means they assign higher probabilities to positive examples than actual probabilities. Moreover, when the model's output probabilities are between 0.7 to 0.9, they approximately behave like a well-calibrated model. In addition to qualitative analysis of calibration curves, they are summarized through brier score calculation. The average brier score of classifiers is summarized in **Table 3**. BS of 0.19 (0.01) for the RF classifier shows its superiority over other classifiers. The result shows the second-best calibrated classifier is GB with BS of 0.20 (0.01). Of six classifiers, SVC, KNN, and MLP together are the worst calibrated classifiers with a BS of 0.25.

In order to evaluate the clinical usefulness of the proposed method, decision curve analysis for 10 independent repeats is performed. As it is depicted in **Figure 6**, net-benefits of almost all classifiers are positive and above the default strategy of intervention-for-non for almost 80% of the threshold range (0.0 to 0.80). Therefore, using the proposed method for decision-making in the mentioned probability range leads to a better outcome than the default strategy of intervention-for-non. Moreover, in the approximate threshold range of 0.70 to 0.9, the RF classifier outperforms other classifiers and stays above both default strategies of intervention-for-all and intervention-for-non. As a result, the outcome of the decisions based on the RF classifier in the mentioned threshold range is superior to both default strategies. In addition, both RF and GB classifiers stay on top of other classifiers and provide better outcomes than other classifiers.

In a 10-fold cross-validation, our second evaluation strategy, the proposed method is evaluated, and the results are summarized in **Table 4**. Similar to the train/test splitting evaluation strategy, the best-performing classifier is RF, with an AUC-ROC of 0.63; the second-best classifier, GB, achieved an AUC-ROC of 0.61. Figure 7 (left) demonstrates the ROC plot of all assessed classifiers. Furthermore, Figure 7 (right) demonstrates that both RF and GB classifiers are significantly different from other classifiers in terms of AUC-ROC.

DISCUSSION

In this study, we evaluated performance of machine learning models for prediction of urolithiasis recurrences using clinical data of patients, demographics, and CT findings. After the comparison of six machine learning models via two evaluation strategy, we revealed that the best ML model is RF, followed by GB, in terms of the overall ROC-AUC, accuracy, and reliability. Also, both RF and GB are originated from a same family of tree-based models, which explains the reported finding.

Both GB and RF are powerful machine learning techniques that are commonly used for predictive modeling. RF, as an ensemble learning technique, constructs numerous decision trees in the training phase. Each tree undergoes training on a randomized portion of the dataset and features, thereby mitigating overfitting and enhancing generalization. In the prediction phase, every tree within the forest autonomously generates predictions, and the ultimate prediction is derived from a majority of individual tree predictions. Random forests exhibit resilience against overfitting and noisy data while demonstrating proficiency in managing a variety of input features[26,27]. GB, an alternative ensemble learning method, constructs a robust predictive model through amalgamating forecasts from numerous weak learners, often in the form of decision trees. Differing from random forests, gradient boosting sequentially develops trees, with each subsequent tree rectifying errors from its predecessors. Every newly added tree aims to diminish residual errors inherited from prior trees. Gradient boosting typically outperforms random forests in predictive accuracy, particularly for structured or tabular data[28].

As sensitivity shown, RF as best performing predictor of urolithiasis recurrence, is able to correctly identify 87% (for train/test splitting strategy) and 90% (for 10-fold cross-validation strategy) of patients at risk. Among identified patients at risk of recurrence by RF, 84% (for train/test splitting strategy) and 83% (for 10-fold cross-validation strategy) actually experience recurrence, as shown by PPV. In other words, RF can predict urolithiasis recurrence with fewer false positives. Besides strong ability of RF to precisely identify patients at risk of recurrence due to its generalizability, RF is also a highly reliable predictor, as evidenced by a Brier score of 0.19 (for train/test splitting strategy) and 0.18 (for 10-fold cross-validation strategy). As it is demonstrated by DCA, prediction of recurrence by using RF predictor is clinically useful and more concretely better than default status of intervention-for-none.

There are also other studies that attempts to predict urolithiasis recurrence using ML models, that mainly relies on the patient's urine biochemistry analysis. In a study by Doyle et al.(29), the feasibility of using 24-

hour urine data of 1231 patients for the prediction of urolithiasis recurrence with ML models has been assessed. They used multiple models, such as RF, GB, LASSO, which revealed that the LASSO model had the best performance in terms of AUC-ROC of 0.62 and 0.63, at two- and five-year recurrence, respectively. On the other hand, in a study by Geraghty et al.[30], the urinary biochemistry measurements, such as pH, volume, urate, oxalate, and calcium, were used to predict stone recurrence using ML models. Although they assessed 7 ML models, none of the utilized models were able to predict the stone recurrence accurately. Another study by Shee et al.[31] which utilizes ML models for predicting urolithiasis recurrence using 24H urine test data of 595 (number of training data = 423 and number of test data = 172) patients, reaches the AUC-ROC of 0.64 on the validation data. Although they performed as same as our RF model in terms of AUC-ROC, our dataset volume is seven times greater than their dataset. Additionally, while they employed the holdout splitting method to partition the data into training and testing sets, we utilized random subsampling with ten repetitions. Therefore, our results are more reliable in comparison with their results.

Several traditional scoring systems, such as S.T.O.N.E., Guy's, and the Seoul National University Renal Stone score, have been developed to assess the complexity of kidney stones. These classification methods aim to provide a standardized approach for evaluating stone characteristics, which can help guide treatment decisions. However, research has indicated that these scoring systems do not show a significant correlation with stone recurrence, suggesting limitations in their predictive power[5]. Therefore, utilizing artificial intelligence methods allows for better prediction of recurrence and management of disease.

While our study shows very promising results, it has some limitations. Although the used dataset is relatively large in comparison with other similar studies, it belongs to a single center. Single-center evaluation of the proposed method may limit the generalizability of the proposed method. Moreover, in the dataset, the number of patients not experiencing urolithiasis recurrence is small relative to the number of patients who experienced urolithiasis recurrence. This imbalance presents challenges in the evaluation process, potentially affecting the accuracy and reliability of predictive analyses. The present study demonstrates that ML-based methods are effective for predicting urolithiasis recurrence with acceptable accuracy compared to traditional scoring systems[5]. This has several implications for clinical practices and future research. Using the proposed method allows clinicians to offer more personalized care. In other words, early identification of patients at high risk of urolithiasis recurrence allows better and cheaper treatment options and improves outcomes.

This study, as the first to utilize CT findings for predicting urolithiasis recurrence, suggests some directions for future research. Future studies could explore the integration of biochemical markers such as blood and 24H urine analysis data, genetic markers, and metabolic data to improve prediction performance, as they are linked to urolithiasis recurrence[32,33]. Additionally, evaluating urolithiasis recurrence prediction methods with external validation data improves generalizability of the prediction methods which is suggested for future works.

CONCLUSIONS

This study aimed to propose an ML-based method for predicting urolithiasis recurrence. Among assessed prediction models, RF performed better than other classifiers with the evidence of sensitivity, PPV, AUC-ROC, and Brier score of 0.87, 0.84, 0.64, 0.19, respectively by train/test splitting strategy and sensitivity, PPV, AUC-ROC, and Brier score of 0.90, 0.83, 0.63, 0.18, respectively by 10-fold cross-validation strategy. The proposed method is the first study to use CT findings in addition to clinical data and baseline characteristics for predicting urolithiasis recurrence, thereby augmenting clinicians' ability to identify high-risk patients and offer more personalized treatment plans. Future research will incorporate extra data to enhance prediction accuracy and employ multi-center datasets for a more comprehensive evaluation of prediction techniques.

REFERENCES

1. Kachkoul R, Touimi GB, El Mouhri G, El Habbani R, Mohim M, Lahrichi A. Urolithiasis: history, epidemiology, aetiologic factors and management. Malays J Pathol. 2023;45:333-52.
2. Banov P, Ceban E. The efficacy of metaphylaxis in treatment of recurrent urolithiasis. J Med Life. 2017;10:188-93.
3. Corbo J, Wang J. Kidney and ureteral stones. Emerg Med Clin North Am. 2019;37:637-48.
4. Baowaidan F, Zugail AS, Lyoubi Y, et al. Incidence and risk factors for urolithiasis recurrence after endourological management of kidney stones: a retrospective single-centre study. Prog Urol. 2022;32:601-7.
5. Chen YH, Li WM, Juan YS, Huang TY, Wang YC, Lee HY. A comparison of S.T.O.N.E nephrolithometry scoring system, Guy's stone score, and Seoul National University Renal Stone Complexity (S-ReSC) in predicting mini-PCNL stone-free rate. Urolithiasis. 2024;52:1-8.
6. Hameed BMZ, Dhavileswarapu AVL S, Raza SZ, et al. Artificial intelligence and its impact on urological diseases and management: a comprehensive review of the literature. J Clin Med. 2021;10:1864.
7. Shah M, Naik N, Somani BK, Hameed BZ. Artificial intelligence (AI) in urology—current use and future directions: an iTRUE study. Turk J Urol. 2020;46:27-39.
8. Homayoun H, Saligheh Rad H, Abbasian Ardakani A. The role of artificial intelligence in urology practice. Transl Res Urol. 2022;4:1-3.
9. Suarez-Ibarrola R, Hein S, Reis G, Gratzke C, Miernik A. Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World J Urol. 2020;38:2329-47.
10. Soleimani Neysiani B, Homayoun H. Medical text and image processing: applications, methods, issues, and challenges. In: Machine Learning and Deep Learning in Medical Data Analytics and Healthcare Applications. 1st ed. CRC Press; 2022.

11. Mitchell T. Machine Learning. McGraw Hill; 1997.
12. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520-5.
13. Theodoridis S, Koutroumbas K. Pattern Recognition. Elsevier; 2009.
14. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. IEEE; 2008:1322-8.
15. Zhang H. The optimality of naive Bayes. AAAI. 2004.
16. Bennett KP, Campbell C. Support vector machines. ACM SIGKDD Explor Newsl. 2000;2:1-13.
17. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
18. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189-232.
19. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13:21-7.
20. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323.
21. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl. 2022;44:875-6.
22. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. 2005:625-32.
23. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861-74.
24. Austin PC, Harrell FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med. 2020;39:2714-42.
25. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
26. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. Int J Comput Sci Issues. 2012;9.
27. Gonzalez R, Saha A, Campbell CJV, Nejat P, Lokker C, Norgan AP. Seeing the random forest through the decision trees. Supporting learning health systems from histopathology with machine learning models: challenges and opportunities. J Pathol Inform. 2024;15:100347.
28. Feng J, Yu Y, Zhou ZH. Multi-layered gradient boosting decision trees. Available from: www.kaggle.com
29. Doyle P, Gong W, Hsi R, Kavoussi N. Machine learning models to predict kidney stone recurrence using 24-hour urine testing and electronic health record-derived features. Res Sq. 2023.
30. Geraghty RM, Wilson I, Olinger E, Cook P, Troup S, et al. Routine urinary biochemistry does not accurately predict stone type nor recurrence in kidney stone formers: a multicentre, multimodel, externally validated machine-learning study. J Endourol. 2023;37:1295-304.
31. Shee K, Liu AW, Chan C, Yang H, Sui W, Desai M, et al. A novel machine-learning algorithm to predict stone recurrence with 24-hour urine data. J Endourol. 2024;38:809-16.
32. Rodjani A, Putra CN, Raharja PAR, et al. Effectivity and safety profile of oxalate decarboxylase in hyperoxaluria patient: a meta-analysis and systematic review. Bali Med J. 2024;13:782-9.
33. Atmoko W, Raharja PAR, Birowo P, et al. Genetic polymorphisms as prognostic factors for recurrent kidney stones: a systematic review and meta-analysis. 2021;16.