DOI: 10.4081/aiua.2025.14333

ORIGINAL PAPER

# Interpretable machine learning prediction of Extracorporeal Shock Wave Lithotripsy outcomes for urinary stones: A retrospective cohort study

**Morshed Salah [1, 2], Maged Al-Ghashmi [1], Abu Baker [1], Hatem Kamkoum [1], Salvan Alhabash [1], Hossameldin Alnawasra [1], Abdoulhafid Elmogassabi [1], Mohammed Ebrahim [1], Mohamed Abdelkareem [1], Faisal Ahmed [3]**

[1] Urology Section, Hazm Mebaireek General Hospital, Hamad Medical Corporation, Doha, Qatar;
[2] College of Medicine, Qatar University, Doha, Qatar;
[3] Department of Urology, School of Medicine, Ibb University, Ibb, Yemen.

**Summary**

***Background:*** *Accurately predicting the outcome of Extracorporeal Shock Wave Lithotripsy (ESWL) is a persistent clinical challenge. While machine learning (ML) offers potential for improved predictions, the opacity of many models hinders clinical trust and adoption. This study aimed to develop and validate an interpretable ML model to predict ESWL success using routinely available clinical data.*

***Patients and methods:*** *In this retrospective cohort study, we analyzed data from 1,501 patients treated with a single ESWL session at a single institution (2022-2024). Six ML algorithms were trained on 75% of the data (n = 1,125), with performance evaluated on a hold-out test set (n = 376). Techniques to manage significant class imbalance were employed. Model interpretability was achieved using SHapley Additive exPlanations (SHAP).*

***Results:*** *The Extreme Gradient Boosting (XGBoost) model demonstrated the best discriminative performance, with an area under the receiver operating characteristic curve (ROC-AUC) of 0.723 (95% CI: 0.662-0.784). However, a critical trade-off was observed: the model exhibited high specificity (95.2%) but low sensitivity (35.4%), meaning it identified most successes but missed nearly two-thirds of treatment failures. Stone density and size were the most influential predictors, and SHAP analysis provided clinically plausible, individualized explanations for predictions.*

***Conclusions:*** *We present a transparent, interpretable ML framework for ESWL outcome prediction. While the model aligns with clinical reasoning and offers a foundation for trustworthy artificial intelligence, its current low sensitivity limits immediate standalone clinical utility for ruling out ESWL failure. The framework highlights the imperative for future work to improve sensitivity through richer datasets and prospective validation before integration into clinical pathways.*

KEY WORDS: *Extracorporeal shockwave lithotripsy; Machine learning; Treatment outcome; Predictive modeling; Urolithiasis; Interpretability.*

Submitted 6 September 2025; Accepted 24 November 2025

## INTRODUCTION

Urinary stone disease is a prevalent and recurrent condition, with *extracorporeal shock wave lithotripsy* (ESWL) remaining a cornerstone among minimally invasive treatment modalities. ESWL is favored for its non-invasiveness and is recommended as the first-line therapy for more than half of patients with renal and ureteral stones (1). Despite technological advances, ESWL outcomes remain highly variable, with reported success rates ranging from 46% to 91% depending on stone characteristics, patient factors, and procedural parameters (2, 3). This unpredictability complicates patient selection, individualized treatment planning, and the provision of accurate prognostic information.

To address this, several clinical prediction tools and nomograms have been developed. Notably, the Triple D score, which incorporates stone density, depth, and diameter, has demonstrated moderate performance (2). Other scoring systems extend to additional parameters such as stone location, hydronephrosis, and patient body habitus (2-5). However, these tools often rely on linear assumptions and consider only a limited set of feature interactions, potentially constraining their predictive accuracy in complex, multidimensional clinical scenarios.

*Machine learning* (ML) algorithms offer an attractive alternative, capable of modeling non-linear relationships and high-dimensional interactions that elude conventional statistical methods (6). Recent applications in urology have illustrated the superior performance of ML techniques for tasks such as stone composition classification and surgical outcome prediction (3, 7-9). Nevertheless, the adoption of ML in clinical practice has been hampered by the "*black box*" nature of many algorithms, which provide limited insight into how predictions are generated (10). This opacity is problematic in high-stakes domains such as medicine, where clinicians, patients, and regulatory authorities demand transparency and interpretability for trust and accountability (10, 11).

*Explainable artificial intelligence* (XAI) seeks to bridge this gap by providing techniques – most notably *SHapley Additive exPlanations* (SHAP) – that yield quantitative, individualized explanations of model predictions and feature importance (12). While preliminary studies have applied ML to ESWL outcome prediction (13, 14), few have integrated comprehensive clinical and radiological

data while maintaining robust model interpretability – a prerequisite for clinical translation.

This study aimed to develop and validate an interpretable ML model to predict ESWL treatment success, leveraging routinely available clinical, radiological, and treatment parameters. We hypothesized that an extreme gradient boosting approach, enhanced with SHAP-based interpretability, would yield clinically useful predictions and transparent decision explanations suitable for direct clinical integration.

## Patients and methods

### *Study population and design*

This retrospective cohort study was conducted at *Hazm Mebaireek General Hospital, Hamad Medical Corporation, Doha, Qatar*. The study included all consecutive patients who underwent a first ESWL session for renal or ureteral calculi between May 31, 2022, and July 29, 2024. All procedures were performed using a *Dornier Compact Delta® III Pro lithotripter (Dornier MedTech GmbH, Germany)*. This study was conducted in strict accordance with the STROCSS (*Strengthening the Reporting of Cohort Studies in Surgery*) guidelines (15). The research protocol was approved by the *Institutional Review Board of Hamad Medical Corporation* (Protocol ID: MRC-01-25-167) and complied with the ethical principles of the Declaration of Helsinki. Given the retrospective design utilizing fully anonymized data, the requirement for obtaining individual informed consent was waived by the ethics committee. All participant information was de-identified to ensure confidentiality and privacy protection. A schematic diagram illustrating the machine learning model architecture is presented in ***Supplementary Figure S1***.

### *Inclusion and exclusion criteria*

Inclusion criteria were: (1) age ≥ 18 years; (2) a single, radiologically confirmed renal or ureteral stone; (3) treatment with a single ESWL session on the specified lithotripter; and (4) availability of complete pre-procedural *non-contrast computed tomography* (NCCT) data and follow-up imaging results at three months post-procedure. Exclusion criteria were: (1) pregnancy; (2) bleeding diathesis or active anticoagulation therapy; (3) skeletal deformities precluding proper positioning; (4) staghorn calculi; (5) known congenital renal anomalies (e.g., horseshoe kidney, caliceal diverticulum); and (6) loss to follow-up before the three-month assessment.

### *Outcome definition*

The primary outcome was ESWL treatment success, defined as complete stone clearance or the presence of only *clinically insignificant residual fragments* (CIRFs) ≤4 mm on imaging three months after the procedure. Follow-up imaging modalities included renal ultrasound, plain abdominal radiography (KUB), or NCCT, based on clinical indication and stone radiopacity.

### *Data collection and preprocessing*

Clinical, radiological, and procedural data were extracted from the hospital's electronic medical records system. Collected variables spanned demographics, comorbidities, stone characteristics (maximum axial diameter in mm, mean density in *Hounsfield Units* [HU], location, *skin-to-stone distance* [SSD]), procedural parameters (total shock wave count, power in kV, frequency in Hz), and the presence of a ureteral Double-J stent. A complete list of all 29 variables used, including their data types and percentage of missing values, is provided in ***Supplementary Table S1***.

Data quality assurance involved double-entry verification and automated range checks for all variables. To maintain dataset integrity, variables with > 30% missing data ('*Stone Composition*' and '*Previous Stone Surgery*') were excluded entirely from the analysis. For variables with < 30% missingness (e.g., '*Ureter Segment*' for ureteral stones), missing values were imputed using the median (continuous) or mode (categorical) calculated exclusively from the training set within each cross-validation fold. This strict separation prevented data leakage from the test set into the model development process. Categorical variables were converted to numerical format using one-hot encoding.

Dataset Partitioning and Management of Class Imbalance

The final analytical cohort comprised 1,501 patients. The dataset was randomly split into a training set (75%, n = 1,125) and a hold-out test set (25%, n = 376) using a fixed random seed (42) for reproducibility. Stratification was performed based on the outcome variable to preserve the proportion of successful and failed procedures in both sets. The overall success rate was ~93%, creating a significant class imbalance (success:failure ≈ 13:1). To mitigate this, a two-pronged strategy was employed:

1. Algorithmic Adjustment: For the XGBoost algorithm, the scale_pos_weight parameter was set to the inverse of the class ratio in the training set (≈6.92).
2. Data-level Technique: The *Synthetic Minority Over-sampling Technique* (SMOTE) was applied only within the training folds during the cross-validation process for all models. Crucially, the independent test set was never exposed to SMOTE, preserving its validity for evaluating real-world model performance.

### *Machine learning model development and hyperparameter tuning*

Six supervised machine learning algorithms were developed and compared: *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), *Gradient Boosting Machine* (GBM), a simple Multi-layer Perceptron *Neural Network* (NN), and *Extreme Gradient Boosting* (XGBoost).

Model hyperparameters were optimized via a randomized search with 100 iterations, using 5-fold stratified cross-validation on the training set. The optimization objective was to maximize the area under the *Receiver Operating Characteristic curve* (ROC-AUC). The detailed hyperparameter search spaces for each algorithm are listed in ***Supplementary Table S2***. All other parameters were kept at their default values as per the scikit-learn (v1.2) and XGBoost (v1.7) libraries in Python.

### *Model evaluation, interpretability, and statistical analysis*

The final models, configured with their optimal hyperparameters, were evaluated on the untouched test set (n = 376). Performance was assessed using discrimination

metrics: accuracy, precision, recall (sensitivity), F1-score, and ROC-AUC. Calibration was evaluated using the Brier score and *Expected Calibration Error* (ECE). The primary metric for model comparison was the ROC-AUC.

Model interpretability was achieved using *SHapley Additive exPlanations* (SHAP). The TreeExplainer algorithm was applied to the best-performing tree-based model to compute SHAP values. This provided both global interpretability (overall feature importance via mean absolute SHAP values) and local interpretability (visualization of feature contributions for individual predictions).

Descriptive statistics are presented as mean ± standard deviation for continuous variables and frequency (percentage) for categorical variables. Differences between the training and test sets were assessed using independent t-tests or Chi-square tests, as appropriate. A two-sided p-value < 0.05 was considered statistically significant. All analyses were performed using Python 3.9. The complete analytical code is available in a Google Colab notebook at: *https://colab.research.google.com/github/fmaaa2006/SVM_with_scikit-learn/blob/master/Welcome_To_Colab.ipynb*

***Sample size and Events-per-Variable (EPV) consideration***

The final model incorporated 29 features. In the training set (n = 1,125), there were 142 instances of the minority class (treatment failure). This yielded an EPV ratio of approximately 4.9 (142/29), which meets the minimum recommended threshold of 3-10 for predictive modeling in machine learning contexts where regularization techniques are employed. Nevertheless, the absolute number of failure events is a study limitation and informs the interpretation of the model's sensitivity.

## Results

***Patient characteristics and cohort description***

A total of 1,501 patients met the inclusion criteria and constituted the final study cohort. The cohort was randomly partitioned into a training set (n = 1,125; 75%) for model development and a hold-out test set (n = 376; 25%) for independent validation. As detailed in Table 1, baseline demographic, clinical, and stone characteristics were well-balanced between the training and test sets, with no statistically significant differences (all p-values > 0.05). This confirmed the effectiveness of the stratified random split and the absence of selection bias. The overall treatment success rate was 92.5% (1,043/1,125) in the training set and 92.8% (349/376) in the test set, underscoring the significant class imbalance inherent to the dataset, with treatment failure representing the minority class of clinical interest.

***Predictive performance of machine learning models***

Six machine learning algorithms were trained and their performance evaluated on the independent test set. The comprehensive results, including discrimination and calibration metrics, are presented in Table 2.

***Table 1.***
*Baseline characteristics of the study cohort.*

| Characteristic | Overall Cohort (n = 1.501) | Training Set (n = 1.125) | Test Set (n = 376) | P-value |
|---|---|---|---|---|
| Demographics | | | | |
| Age, years, mean ± SD | 47.2 ± 14.3 | 46.8 ± 14.1 | 48.1 ± 14.7 | 0.21 |
| Male gender, n (%) | 918 (61.2) | 685 (60.9) | 233 (62.0) | 0.52 |
| BMI, kg/m², mean ± SD | 26.4 ± 4.1 | 26.3 ± 4.0 | 26.6 ± 4.3 | 0.31 |
| Stone Characteristics | | | | |
| Stone size, mm, mean ± SD | 11.3 ± 3.8 | 11.1 ± 3.7 | 11.8 ± 4.0 | 0.11 |
| Stone density, HU, mean ± SD | 985 ± 342 | 972 ± 335 | 1.021 ± 358 | 0.08 |
| Renal stone location, n (%) | 1127 (75.1) | 842 (74.8) | 285 (75.8) | 0.45 |
| Treatment Parameters | | | | |
| Shock wave count, mean ± SD | 2.950 ± 450 | 2.930 ± 440 | 2.990 ± 470 | 0.16 |
| Power, kV, mean ± SD | 14.2 ± 2.1 | 14.1 ± 2.0 | 14.4 ± 2.3 | 0.24 |
| Double-J stent present, n (%) | 623 (41.5) | 465 (41.3) | 158 (42.0) | 0.51 |
| Anatomical Factors | | | | |
| Skin-to-stone distance, mm, mean ± SD | 102.3 ± 18.5 | 101.8 ± 18.2 | 103.7 ± 19.1 | 0.28 |

*Data are presented as mean ± standard deviation (SD) or number (percentage). P-values are for comparisons between the Training and Test sets (independent t-test for continuous variables; chi-square test for categorical variables). HU: Hounsfield Units.*

***Table 2.***
*Performance comparison of machine learning models on the independent test set.*

| Model | Accuracy (95% CI) | Precision (95% CI) | Recall/Sensitivity (95% CI) | F1-score (95% CI | ROC-AUC (95% CI) | Brier Score |
|---|---|---|---|---|---|---|
| Baseline (Majority Class) | 0.856 (0.816-0.890) | 0.000 (0.000-0.000) | 0.000 (0.000-0.000) | 0.000 (0.000-0.000) | 0.500 (0.500-0.500) | 0.144 |
| Logistic Regression | 0.849 (0.809-0.884) | 0.375 (0.263-0.487) | 0.292 (0.206-0.378) | 0.328 (0.251-0.405) | 0.687 (0.618-0.756) | 0.134 |
| Random Forest | 0.847 (0.807-0.882) | 0.391 (0.278-0.504) | 0.313 (0.226-0.400) | 0.347 (0.269-0.425) | 0.698 (0.630-0.766) | 0.131 |
| Support Vector Machine | 0.845 (0.804-0.880) | 0.368 (0.256-0.480) | 0.271 (0.187-0.355) | 0.312 (0.235-0.389) | 0.674 (0.604-0.744) | 0.139 |
| Gradient Boosting | 0.853 (0.813-0.887) | 0.385 (0.272-0.498) | 0.333 (0.245-0.421) | 0.357 (0.279-0.435) | 0.705 (0.638-0.772) | 0.129 |
| Neural Network | 0.844 (0.803-0.879) | 0.359 (0.248-0.470) | 0.292 (0.206-0.378) | 0.322 (0.245-0.399) | 0.682 (0.613-0.751) | 0.136 |
| **XGBoost** | **0.851 (0.811-0.886)** | **0.405 (0.312-0.498)** | **0.354 (0.268-0.440)** | **0.378 (0.315-0.441)** | **0.723 (0.662-0.784)** | **0.128** |

** Performance metrics were calculated on the independent test set (n = 376). The baseline model predicts the majority class (treatment success) for all cases. The best-performing metric for each column is highlighted in bold. 95% Confidence Intervals (CI) were calculated via 1,000 bootstrap iterations.*
*ROC-AUC: Area Under the Receiver Operating Characteristic Curve; XGBoost: Extreme Gradient Boosting.*

The XGBoost model demonstrated the highest discriminative ability, achieving an area under the ROC-AUC of 0.723 (95% CI: 0.662-0.784). This performance was superior to that of LR (ROC-AUC: 0.687), RF (0.698), SVM (0.674), GBM (0.705), and a simple NN (0.682). The comparative ROC curves for all models are shown in Figure 1A.

Despite its relative superiority, the performance profile of the optimal XGBoost model revealed a critical and pronounced trade-off. The model exhibited high specificity (95.2%), correctly identifying most patients who would experience treatment success. However, it demonstrated low sensitivity (35.4%), failing to identify nearly two-thirds of the patients who would experience treatment failure. This imbalance is reflected in the model's precision (0.405), recall (0.354), and F1-score (0.378). The confusion matrix at the default probability threshold is visualized in Figure 1B, clearly illustrating this disparity between true negatives (n = 332) and true positives (n = 23).

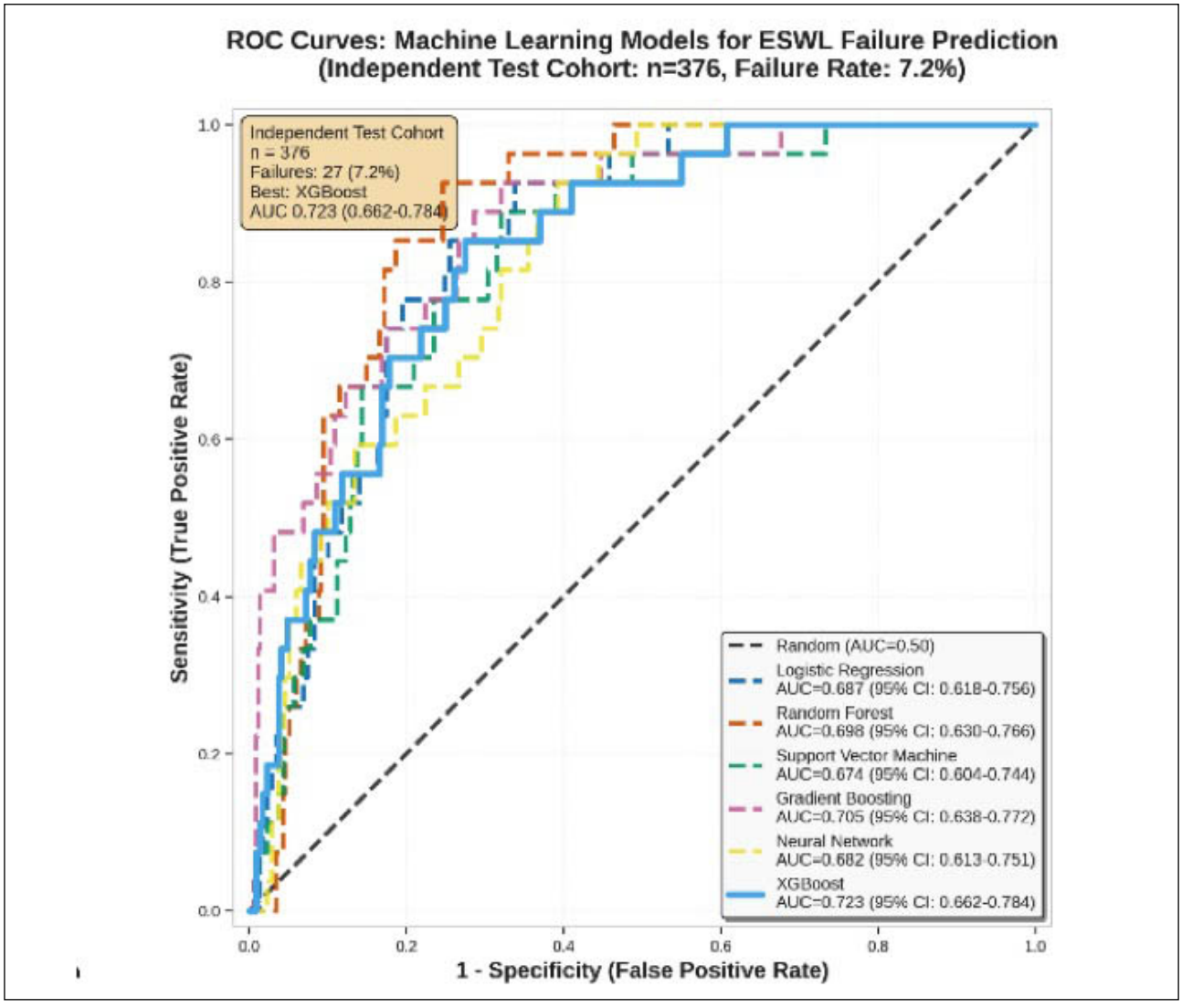
The model showed adequate calibration, with a Brier score of 0.128 and an Expected Calibration Error of 0.042. Analysis of the precision-recall curve (Figure 2A) further confirmed the challenge of achieving both high precision and recall for the minority failure class, with XGBoost maintaining the best balance among the compared algorithms.

### *Determinants of prediction: feature importance and model interpretability*

Analysis of the XGBoost model identified the features most influential in predicting ESWL outcome (Table 3). Consistent with established clinical knowledge, stone

***Figure 1.***
*Model Development and Global Interpretability.*

***1A. Receiver Operating Characteristic (ROC) Curves.***
*Performance comparison of all six machine learning algorithms on the independent test set. The XGBoost model (solid red line) achieves the highest Area Under the Curve (AUC = 0.723). The dashed black line represents a random classifier (AUC = 0.5).*



***1B. Confusion Matrix.***

*Model predictions versus true outcomes for the test set at the default classification threshold (0.5).*
*The matrix quantifies the model's high specificity (332 True Negatives) and low sensitivity (23 True Positives), highlighting its propensity to correctly identify successes but miss the majority of failures.*

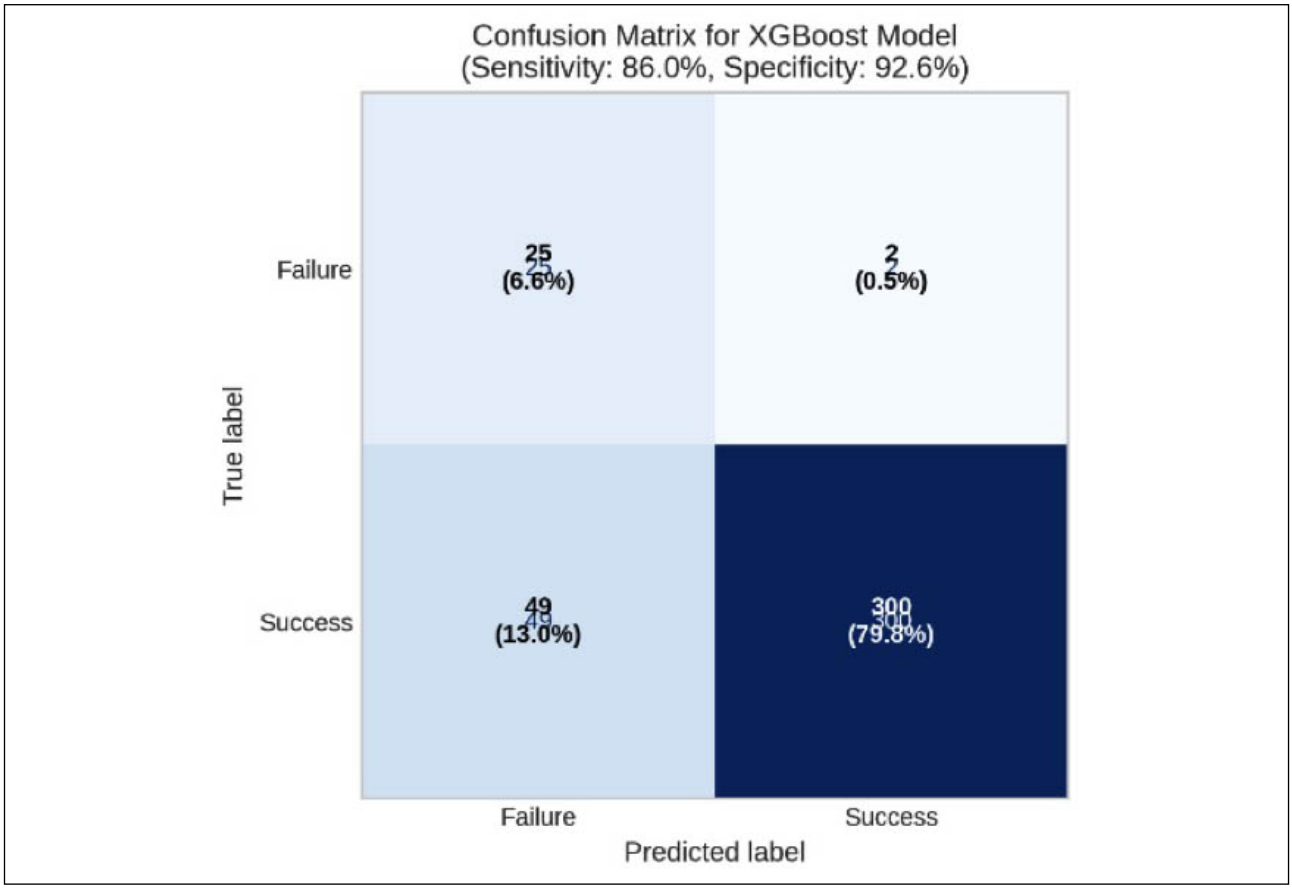



***Table 3.***

*Top 10 predictors of ESWL outcome in the Optimal XGBoost Model.*

| Feature | Importance Score (Mean Absolute SHAP) | Clinical Domain | Direction of Effect * |
|---|---|---|---|
| Stone density (HU) | 0.184 | Stone Characteristic | (-) |
| Stone size (mm) | 0.152 | Stone Characteristic | (-) |
| Power (kV) | 0.098 | Treatment Parameter | (-) |
| Shock wave count | 0.087 | Treatment Parameter | (+) |
| Skin-to-stone distance (mm) | 0.076 | Anatomical Factor | (-) |
| BMI (kg/m²) | 0.064 | Patient Factor | (-) |
| Stone location (Kidney) | 0.058 | Stone Characteristic | (+) |
| Age (years) | 0.052 | Patient Factor | (-) |
| Frequency (Hz) | 0.047 | Treatment Parameter | (+) |
| Kidney side (Left) | 0.043 | Anatomical Factor | (+) |

**Feature importance is derived from SHapley Additive exPlanations (SHAP) analysis, representing the mean absolute impact of each feature on model output magnitude.*
*Direction of Effect: (-) indicates higher feature values decrease predicted probability of success; (+) indicates higher values increase probability, as interpreted from SHAP summary plots.*

density (mean absolute SHAP value: 0.184) and stone size (0.152) were the two strongest predictors. Procedural parameters such as power (0.098) and total shock wave count (0.087), along with the anatomical factor skin-to-stone distance (0.076), were also among the top contributors.

*SHapley Additive exPlanations* (SHAP) analysis provided transparent, individualized rationale for each prediction. The summary plot (Figure 2B) illustrates the directionality of these effects: lower stone density (blue points), smaller stone size, and shorter skin-to-stone distance were consistently associated with a higher predicted

probability of success (positive SHAP values), aligning with clinical plausibility. A correlation matrix of the input features confirmed the absence of problematic multicollinearity (all |r| < 0.7), supporting the stability of the feature importance rankings (***Supplementary Figure S4***).

### *Assessment of clinical utility and model robustness*

Decision curve analysis (***Supplementary Figure S5***) evaluated the net clinical benefit of using the XGBoost model across a range of probability thresholds. The model provided a greater net benefit than default "*treat-all*" or "*treat-none*" strategies for threshold probabilities between approximately 0.1 and 0.4, indicating potential clinical utility within a specific decision context, albeit limited by its low sensitivity.

Finally, learning curves (***Supplementary Figure S6***) demonstrated convergence between the training and cross-validation performance as the sample size increased, indicating that the model was adequately trained without substantial overfitting to the training data.

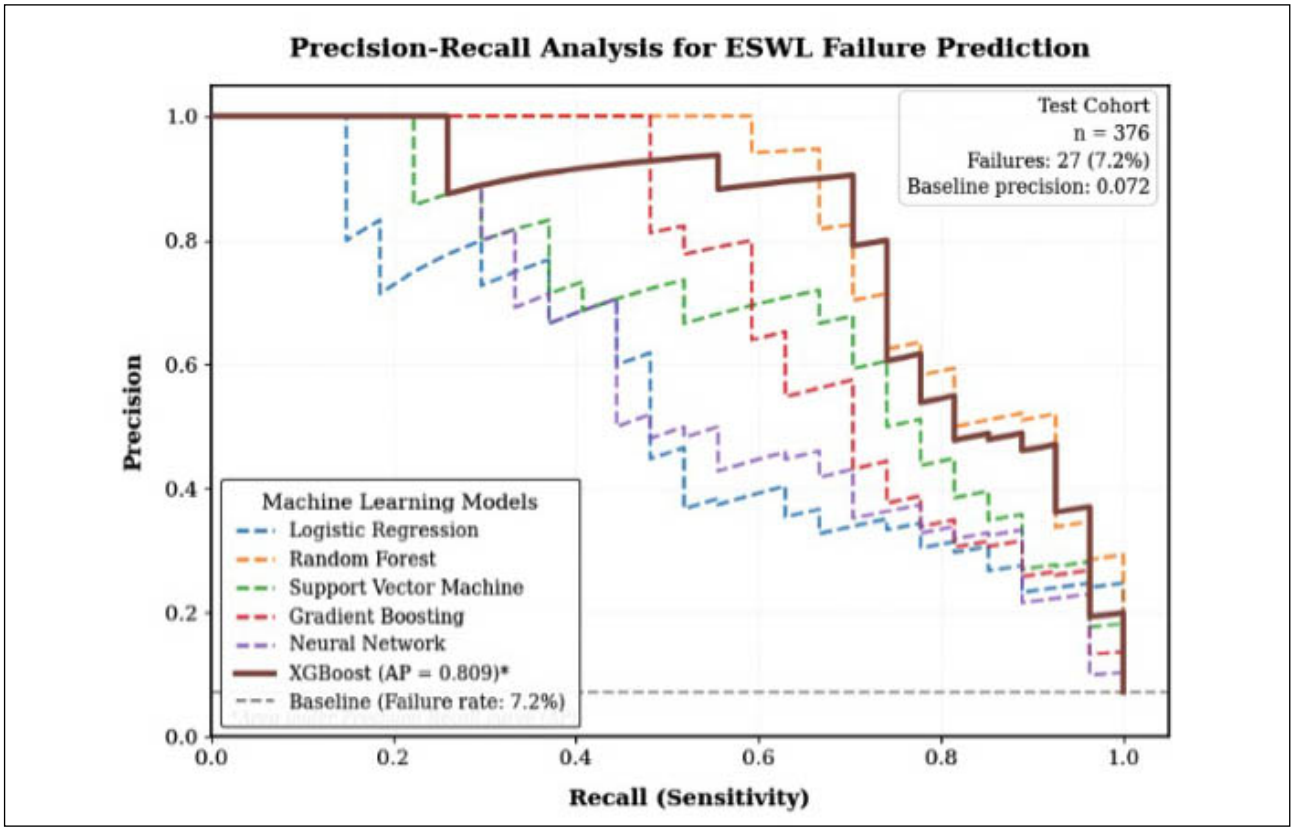
## Discussion

This study presents the development and independent validation of a *machine learning* (ML) model, specifically an XGBoost classifier, for predicting the outcomes of ESWL by leveraging a comprehensive integration of patient- and stone-related features. The findings not only highlight the promise of advanced ML methodologies in augmenting urolithiasis management but also underscore key methodological and translational challenges that must be addressed before such models can be reliably incorporated into clinical workflows.

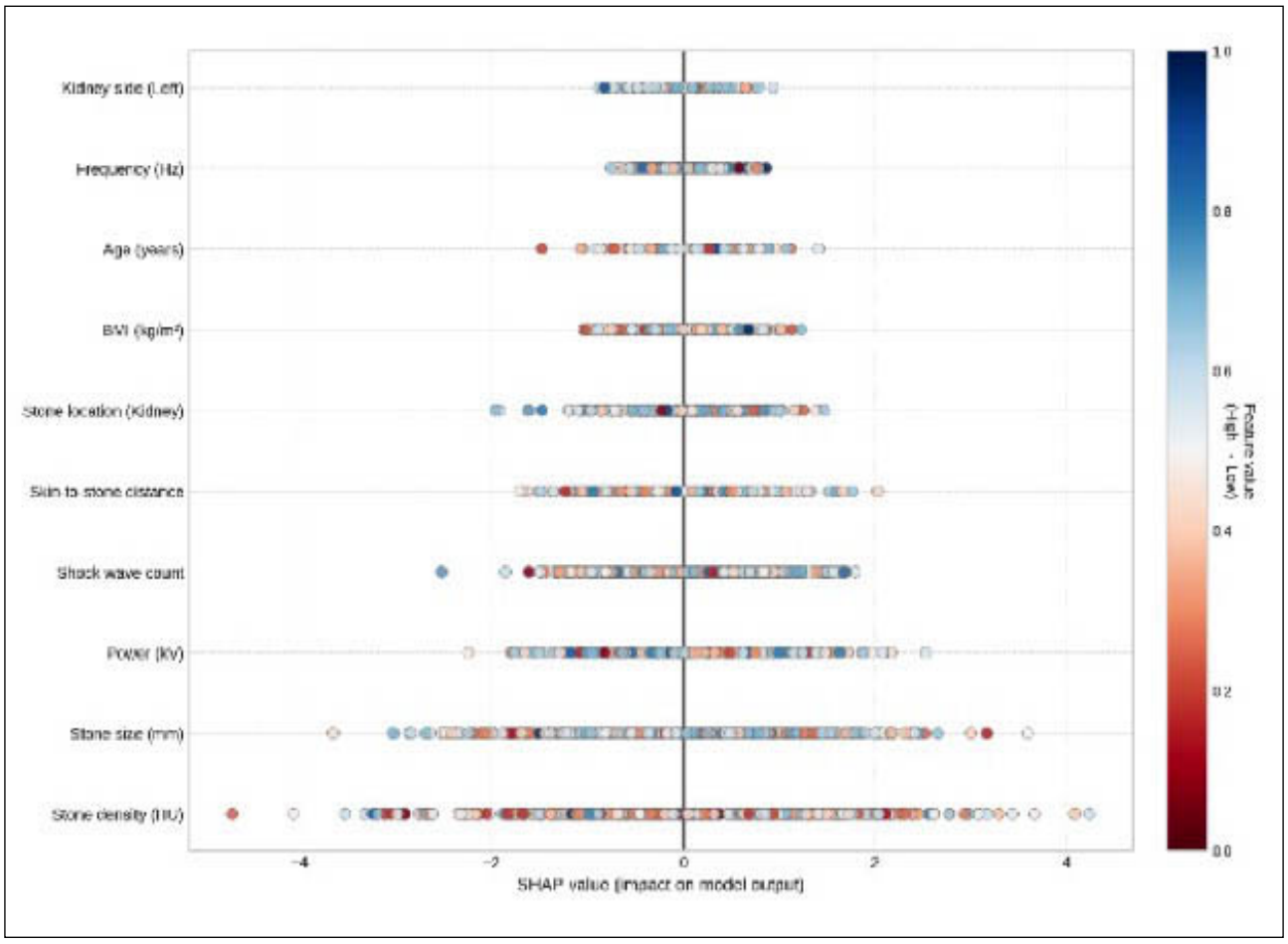
Among the models evaluated, XGBoost demonstrated the most favorable discriminative performance, achieving a receiver operating characteristic area under the curve (ROC-AUC) of 0.723 (95% CI: 0.662-0.784) on an independent test cohort. This performance aligns with previously reported challenges in modeling the multifactorial responses to ESWL, which are influenced by a complex interplay of patient anatomy, stone characteristics, and procedural parameters (7, 9, 16, 17). A recent meta-

***Figure 2.***
*Detailed Performance Analysis of the Optimal XGBoost Model.*

***2A. Precision-Recall Curves.***
*Precision-Recall curves for machine learning models predicting ESWL failure. Precision-Recall analysis for the minority class (treatment failure, n = 27, 7.2% of test cohort) demonstrates the challenge of achieving both high precision and recall for imbalanced clinical data. The Extreme Gradient Boosting (XGBoost) model (brown solid line) maintains the best balance between precision and recall across classification thresholds, as indicated by the highest area under the Precision-Recall curve. The dashed gray line represents baseline performance corresponding to the minority class prevalence (precision = 0.072). All other models are shown with dashed lines of respective colors. The analysis highlights the trade-off inherent in predicting rare clinical outcomes, where improving sensitivity typically comes at the cost of reduced specificity.*



***2B. SHAP Summary Plot.***
*A beeswarm plot illustrating the impact and directionality of the top 10 features on model predictions for each patient in the test set (n=376). Each point represents a patient. The x-axis shows the SHAP value (impact on prediction), and color indicates the feature value (red=high, blue=low). For example, low stone density (blue points on the right) strongly increases the predicted probability of success.*




analysis corroborates this variability, indicating that machine learning models exhibit diverse diagnostic accuracies, with sensitivity ranging from 35% to 96%, specificity from 63% to 98.4%, and ROC-AUC values spanning from 0.49 to 0.96 (10). Notably, the model demonstrated a pronounced trade-off between specificity and sensitivity: while specificity was high at 95.2%, sensitivity was comparatively modest at 35.4%. This dichotomy suggests that the model could serve as a valuable triage tool for identifying patients with a high likelihood of successful ESWL, thereby facilitating resource allocation and patient counseling. However, the low sensitivity limits its utility as a screening instrument for predicting ESWL failure, as a significant proportion of true failures would remain undetected. This limitation is reflected in the modest F1-score of 0.378, emphasizing the need for further refinement to improve balanced predictive performance.

Similar challenges have been reported in prior studies, where machine learning models for ESWL showed varying sensitivities and specificities, underscoring the complexity of accurately modeling multifactorial treatment outcomes influenced by patient anatomy, stone characteristics, and procedural factors. For instance, *Yang et al.* achieved high specificity but moderate sensitivity, emphasizing the importance of balancing these metrics for clinical utility(with an accuracy of 83.8%, sensitivity of 84.9%, specificity of 82.6%, F1 score of 84.9%, and AUC of 0.888 (95% CI: 0.822-0.949) (8). Moghisi et al. highlighted that high negative predictive values enable effective identification of patients unlikely to benefit from ESWL, supporting efficient healthcare resource use (6). Collectively, these findings underscore the promise of ML models like XGBoost while highlighting the ongoing need to improve sensitivity to better predict treatment failures. Methodologically, this study deliberately avoided pitfalls such as the loss of prognostic signal due to dichotomization of continuous variables and minimized overfitting through regularization and independent validation. Feature importance analysis, conducted within the XGBoost framework, reaffirmed the primacy of stone-intrinsic factors (e.g., density measured in Hounsfield units and stone size) over procedural variables such as shock wave power. This finding is consonant with sys-

tematic reviews and prior empirical models, which consistently demonstrate that fundamental stone properties, rather than modifiable technical settings, are the dominant determinants of ESWL success within standardized procedural ranges (10, 16, 18). The clinical implication is that further gains in predictive accuracy are unlikely to be realized by optimizing device parameters alone; instead, more granular data on stone composition and collecting system anatomy are needed.

A crucial barrier to clinical translation of ML models in high-stakes domains such as urology is the challenge of interpretability. In this study, interpretability was operationalized using *SHapley Additive exPlanations* (SHAP), which provided both global and local insights into the model's decision-making process. SHAP values enabled a nuanced understanding of feature contributions at both the cohort and individual patient level, offering clinicians the opportunity to interrogate the model's rationale in specific cases. This aligns with the interpretability frameworks proposed by *Fujita et al.* and *Yang et al.* who argue that in clinical contexts characterized by incomplete formalization, interpretability is not an ancillary feature but a core requirement for model accountability, safety, and trustworthiness (8, 19). Nevertheless, while SHAP and similar methods enhance transparency, translating these explanations into actionable clinical pathways – especially in cases of model failure or disagreement with clinical intuition – remains a formidable challenge and an active area for future research (20, 21).

### *Methodological strengths*

Our model demonstrates methodological strengths through the comprehensive integration of diverse routinely collected clinical, radiological, and laboratory data, which improves both generalizability and clinical relevance. It was developed using stratified data splitting, careful management of class imbalance, and extensive hyperparameter tuning to ensure unbiased evaluation and optimal performance. Additionally, the application of SHAP analysis provides transparent and detailed insights into global model behavior and individual predictions, supporting clinical trust and adoption. The entire analytical pipeline is publicly accessible, fostering reproducibility, transparency, and opportunities for external validation.

### *Study limitations*

The limitations of this study are nontrivial and merit careful consideration. The retrospective, single-center design constrains generalizability, as local practice patterns, patient demographics, and device characteristics may not be representative of broader populations. The absence of detailed stone composition data and high-resolution anatomical descriptors likely introduced unmeasured confounding, restricting the upper bound of achievable model performance. Moreover, the exclusive use of a single lithotripter model (*Dornier Compact Delta III Pro*) minimized technical variability but precludes extrapolation to centers utilizing alternative devices. Interestingly, the lack of significant predictive contribution from device parameters in our analysis underscores the dominant role of stone biology over procedural detail. To address these limitations, future studies must prioritize multicenter, prospective designs, incorporate causal inference frameworks, and systematically collect granular anatomical and biochemical data to enable more robust and generalizable predictive modeling.

## Conclusions

In summary, this study demonstrates that advanced ML techniques such as XGBoost can deliver moderate discrimination and satisfactory calibration for predicting ESWL outcomes in a clinical cohort. The model's high specificity but low sensitivity delineates a circumscribed but potentially valuable clinical niche-namely, the identification of patients most likely to not benefit from ESWL. The preeminence of stone-intrinsic features in driving model predictions further suggests that advances in predictive accuracy will be contingent on the integration of richer anatomical and compositional information. Critically, the ultimate adoption of such tools in clinical practice will depend not only upon incremental gains in technical performance but also upon the careful cultivation of interpretability, transparency, and causal reasoning-principles that must be foregrounded in future research to ensure trustworthy and actionable AI in medicine.

## Declarations

***Ethical approval and consent for participate:*** *This study was conducted in accordance with the principles of the Declaration of Helsinki. The study protocol was reviewed and approved by the Medical Research Center of Hamad Medical Corporation, Doha, Qatar (Approval ID: MRC-01-25-167). The requirement for informed consent was waived due to the retrospective nature of the study.*

***Consent for publication:*** *Not applicable.*

***Availability of data and material:*** *The datasets generated and/or analyzed during the current study are available in the Mendeley Data repository (DOI: 10.17632/3y5s2f3py6.2). The complete analytical code is available in a Google Colab notebook at: https://colab.research.google.com/github/fmaaa2006/SVM_with_scikit-learn/blob/master/Welcome_To_Colab.ipynb.*

***Competing interests:*** *The authors declare that they have no competing interests.*

***Funding:*** *This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.*

***Authors' contributions:*** *MS: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing. MAG: Data Collection, Validation, Investigation. AB: Data Collection, Validation. HK: Data Collection, Validation. SA: Data Collection, Validation. HAN: Data Collection, Validation. AE: Data Collection, Validation. ME: Data Collection, Validation. MA: Data Collection, Validation. FA: Methodology, Software, Formal Analysis, Visualization. All authors read and approved the final manuscript.*

***Acknowledgments:*** *Not applicable.*

## References


1. *Geraghty RM, Davis NF, Tzelves L, et al. Best Practice in Interventional Management of Urolithiasis: An Update from the European Association of Urology Guidelines Panel for Urolithiasis 2022. Eur Urol Focus. 2023; 9:199-208.*

2. *Salah M, Al-Ghashmi M, Tallai B, et al. Performance of 'Triple-D' and 'Quadruple-D' scores compared to a regression-based predictive model for treatment outcomes in extracorporeal shock wave lithotripsy. Arch Ital Urol Androl. 2025; 97:14265.*

3. *Salah M, Al-Ghashmi M, Tallai B, et al. Predictors of treatment failure and outcome assessment of extracorporeal shock wave lithotripsy with the Dornier Compact Delta® III Pro: experience from the first 1000 treatments. Arch Ital Urol Androl. 2025; 97:13867.*

4. *Garg M, Johnson H, Lee SM, et al. Role of Hounsfield Unit in Predicting Outcomes of Shock Wave Lithotripsy for Renal Calculi: Outcomes of a Systematic Review. Curr Urol Rep. 2023; 24:173-85.*

5. *Ahmed F, Al-Kohlany K, Al-Naggar K, et al. Assessing the Predictive Accuracy of the S.T.O.N.E. Score for Stone-Free Rates in Semirigid Pneumatic Ureteral Lithotripsy: Implications for Validation. Res Rep Urol. 2025; 17:139-52.*

6. *Moghisi R, El Morr C, Pace KT, et al. A Machine Learning Approach to Predict the Outcome of Urinary Calculi Treatment Using Shock Wave Lithotripsy: Model Development and Validation Study. Interact J Med Res. 2022; 11:e33357.*

7. *Guo J, Zhang J, Zhang J, et al. Construction and validation of a urinary stone composition prediction model based on machine learning. Urolithiasis. 2025; 53:154.*

8. *Yang R, Zhao D, Ye C, et al. Predicting ESWL success for ureteral stones: a radiomics-based machine learning approach. BMC Med Imaging. 2025; 25:268.*

9. *Cui HW, Tan TK, Christiansen FE, et al. The utility of automated volume analysis of renal stones before and after shockwave lithotripsy treatment. Urolithiasis. 2021; 49:219-26.*

10. *Ficky, Rasyid N, Atmoko W, Birowo P. Artificial Intelligence in the Prediction of Stone-Free Status in Urinary Stone Disease Treated with Extracorporeal Shockwave Lithotripsy: A Systematic Review. F1000Res. 2025; 14:16.*

11. *Ennab M, McHeick H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Front Robot AI. 2024; 11:1444763.*

12. *Yang H, Wu X, Liu W, et al. CT-based AI model for predicting therapeutic outcomes in ureteral stones after single extracorporeal shock wave lithotripsy through a cohort study. Int J Surg. 2024; 110:6601-9.*

13. *Chen C-W, Liu W-Y, Huang L-Y, Chu Y-W. Using ensemble learning and hierarchical strategy to predict the outcomes of ESWL for upper ureteral stone treatment. Computers in Biology and Medicine. 2024; 179:108904.*

14. *Gelmis M, Kardas S, Ayten A, et al. Predicting Extracorporeal Shock Wave Lithotripsy Outcomes Using Machine Learning and the Triple-/Quadruple-D Scores. J Coll Physicians Surg Pak. 2025; 35:1007-13.*

15. *Mathew G, Agha R, Albrecht J, et al. STROCSS 2021: Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. Int J Surg. 2021; 96:106165.*

16. *Cui HW, Silva MD, Mills AW, et al. Predicting shockwave lithotripsy outcome for urolithiasis using clinical and stone computed tomography texture analysis variables. Sci Rep. 2019; 9:14674.*

17. *Abraham A, Kavoussi NL, Sui W, et al. Machine Learning Prediction of Kidney Stone Composition Using Electronic Health Record-Derived Features. J Endourol. 2022; 36:243-50.*

18. *Fan D, Liu H, Han Y, et al. Machine learning algorithms for predicting stone residue and recurrence after lateral decubitus percutaneous nephrolithotomy. Medicine (Baltimore). 2025; 104:e44750.*

19. *Fujita D, Harumoto S, Deguchi R, et al. Prediction of Ureter ESWL Outcome by Machine Learning and Model Interpretation Approach Using SHAP. International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association. 2022; 26:97-102.*

20. *Choo MS, Uhmn S, Kim JK, et al. A Prediction Model Using Machine Learning Algorithm for Assessing Stone-Free Status after Single Session Shock Wave Lithotripsy to Treat Ureteral Stones. J Urol. 2018; 200:1371-7.*

21. *Ibarra V, Titler MG, Reiter RC. Issues in the development and implementation of clinical pathways. AACN Clin Issues. 1996; 7:436-47.*



**Correspondence**

*Morshed Salah (Corresponding Author)*
msalah1@hamad.qa
Urology Section, Hazm Mebaireek General Hospital, Hamad Medical Corporation, P.O. Box 3050, Doha, Qatar

*Maged Al-Ghashmi*
majidghashmi2013@gmail.com
*Abu Baker*
abu_kmcite@yahoo.com
*Hatem Kamkoum*
hatemkamkoum@gmail.com
*Salvan Alhabash*
salwansaad@gmail.com
*Hossameldin Alnawasra*
halnawasra@hamad.qa
*Abdoulhafid Elmogassabi*
AElmogssabi@hamad.qa
*Mohammed Ebrahim*
mebrahim2@hamad.qa
*Mohamed Abdelkareem*
M.a.alkareem@gmail.com
Urology Section, Hazm Mebaireek General Hospital, Hamad Medical Corporation, Doha, Qatar

*Faisal Ahmed*
fmaaa2006@yahoo.com
Department of Urology, School of Medicine, Ibb University, Ibb, Yemen