Power System Technology
ISSN:1000-3673

Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

Comparative Evaluation of Machine Learning Models for Reliable Kidney Stone Prediction: Performance, Robustness, and Clinical Utility

[1]J R Harshavardhan,

Research Scholar, Dept. of CSE, R N S Institute of Technology, Associate Professor, KSSEM, Bangalore-India

[2] Dr. Anjan Kumar K N,

Associate Professor, Dept. of CSE, R N S Institute of Technology, Bengaluru- India

[3] Dr. Satish Kumar S,

Professor, Dept. of ISE, R N S Institute of Technology, Bengaluru- India.

[4] Nandisha A C,

Assistant Professor, Dept. of AI&ML, City Engineering College, Bengaluru- India.

[1]sudhaharshavardhan@gmail.com, [2]anjankn05@gmail.com, [3] sathish_tri@yahoo.com,

[4] nandishac@gmail.com

Abstract: - Kidney stone disease nephrolithiasis represents a growing global health concern, necessitating accurate and early prediction to prevent complications and optimize healthcare resources. This study conducts a comprehensive comparative assessment of multiple supervised machine learning algorithms Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, and Neural Networks for predicting kidney stone formation. Model performance was systematically evaluated using a suite of statistical and clinical metrics to determine predictive accuracy, robustness, and interpretability. The findings indicate that ensemble-based models, particularly Random Forest and Gradient Boosting, consistently outperform other algorithms in terms of accuracy and generalization capability. In contrast, simpler models such as Logistic Regression and Decision Trees offer enhanced interpretability, making them more suitable for clinical decision support. To enhance transparency and clinical trust, SHAP (SHapley Additive exPlanations) analysis was employed to elucidate feature contributions, identifying serum calcium and uric acid as the most influential biomarkers in prediction outcomes. The study underscores the critical balance between predictive performance, model robustness, and interpretability, demonstrating that explainable and validated machine learning frameworks can serve as effective tools for early risk stratification, timely clinical intervention, and improved patient management in nephrolithiasis care.

Keywords: Kidney stone prediction, Nephrolithiasis, Machine learning, Comparative evaluation, Predictive modeling, Clinical decision support, Model interpretability, Random Forest, Gradient Boosting, Logistic Regression, Robustness, Healthcare analytics.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

1. Introduction

Nephrolithiasis, commonly known as kidney stone disease, is among the most prevalent urological conditions and affects millions of individuals globally. Over recent decades, its occurrence has shown a consistent upward trend primarily attributed to dietary changes, sedentary habits, and an increased prevalence of metabolic disorders. Epidemiological data indicate a global lifetime risk of 10–15%, with recurrence rates approaching 50% within 5–10 years of the first episode [1]. Apart from causing intense pain and discomfort, this condition also places significant financial strain and increases the likelihood of developing long-term complications, including chronic kidney disease [5]. These challenges underscore the importance of early detection and accurate risk prediction to improve patient outcomes and reduce healthcare costs.

Conventional diagnostic practices, including ultrasound, CT scans, X-rays, and biochemical tests, remain effective but are resource-intensive, costly, and limited in their predictive capacity, particularly before symptom onset [7]. With the expansion of electronic health records and clinical datasets, coupled with advances in computational methods, machine learning (ML) offers new opportunities for predictive modeling and preventive care in nephrolithiasis.

Machine learning techniques are highly effective in uncovering complex, non-linear associations among demographic, clinical, and biochemical variables that conventional statistical approaches often fail to detect [3]. However, comparative evaluations of different ML approaches remain limited, particularly with respect to predictive accuracy, robustness across diverse datasets, and interpretability for clinical application [6]. In healthcare, model transparency is as critical as accuracy, since explainability fosters clinician trust and adoption.

This study aimed to bridge these gaps by systematically assessing six machine learning algorithms logistic regression, decision trees, random forests, support vector machines, gradient boosting, and neural networks using a carefully curated kidney stone dataset [10]. The models were assessed not only for predictive accuracy but also for robustness through cross-validation and sensitivity analysis, as well as interpretability using feature-importance and explainability methods.

By combining model performance, stability, and clinical applicability, this research strengthens the expanding body of evidence on machine learning applications in nephrology. The outcomes underscore the inherent trade-offs between predictive accuracy and interpretability, providing valuable guidance for the implementation of ML-driven decision support systems aimed at kidney stone risk prediction, prevention, and clinical management.

2.Literature Review

Machine learning models were developed and validated to predict kidney stone recurrence by integrating 24-hour urinary chemistry parameters with electronic health record (EHR) data. A

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

range of algorithms including logistic regression, random forest, and gradient boosting have been implemented, with ensemble approaches delivering superior predictive accuracy compared to traditional diagnostic methods. The findings revealed that biochemical indicators such as calcium, oxalate, and uric acid, when combined with demographic and clinical attributes, substantially improved the predictive performance and strengthened the clinical decision-support potential of the proposed framework. These insights illustrate the value of ML-based tools in tailoring recurrence risk assessments, enabling more proactive follow-up and individualized management of patients with nephrolithiasis [1].

A clinical nomogram was developed to estimate the likelihood of kidney stone recurrence by integrating patient demographics, medical history, and stone-related risk factors. Utilizing data from a large patient cohort, the study applied statistical modeling to identify key predictors and convert them into an easy-to-use scoring framework. The resulting tool demonstrated high discriminative accuracy and excellent calibration, providing clinicians with a robust, individualized approach for assessing the recurrence risk in patients with nephrolithiasis. Thus, it provides an early framework for personalized nephrolithiasis management, supporting targeted monitoring and preventive care before the adoption of machine learning based models [2].

This highlights the need for a predictive tool for symptomatic kidney stone episodes as an initial step toward advancing personalized care in nephrolithiasis. This study emphasizes the need for dependable risk stratification methods that go beyond generalized population statistics to deliver individualized forecasts of recurrence and symptoms. By incorporating patient-specific variables into a predictive framework, this tool provides a basis for tailoring follow-up and prevention strategies, and therapeutic decisions. This commentary underscores the emerging role of predictive analytics in urology, paving the way for more sophisticated, machine learning based approaches in the future [3].

The use of machine learning models to predict kidney stone composition from electronic health record (EHR)-derived data. By incorporating demographic, clinical, and laboratory information, the models were trained to distinguish between clinically important stone types, such as uric acid and non-uric acid stones. Ensemble approaches achieved the highest predictive accuracy, whereas feature importance analysis demonstrated that routinely available EHR variables can act as effective, non-invasive predictors. The study underscores the promise of ML in providing cost-effective, non-invasive decision support for nephrolithiasis management, facilitating personalized treatment selection and preventive care strategies [4].

This study employed machine learning techniques to predict kidney stone composition using clinical and laboratory parameters derived from electronic health records (EHRs). Multiple predictive models were developed and evaluated, with ensemble approaches achieving the highest accuracy in distinguishing uric acid stones from other types. These results indicate that easily accessible EHR data can be effectively utilized for stone type prediction, thereby minimizing the need for invasive or expensive diagnostic methods. Overall, this study

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

illustrates the potential of ML to enable non-invasive, data-driven decision support and foster personalized strategies for kidney stone management [5].

This study introduced an AI-driven machine learning framework developed to predict renal stone recurrence following Endoscopic Combined Intrarenal Surgery (ECIRS). The proposed model integrates 24-hour urinary chemistry data with postoperative clinical parameters to effectively stratify patients according to recurrence risk, demonstrating superior accuracy compared to conventional approaches. Beyond its predictive performance, the framework prioritizes clinical interpretability, highlighting urinary biochemical factors as key determinants of stone recurrence. These findings highlight the role of ML-based risk stratification tools in guiding personalized follow-up care and preventive strategies for patients undergoing ECIRS [6].

This study presents an AI-based machine learning model designed to predict renal stone recurrence after Endoscopic Combined Intrarenal Surgery (ECIRS). By integrating 24-hour urinary profiles with critical clinical parameters, the authors developed a postoperative risk stratification framework that achieved high predictive accuracy. The analysis identified urinary and metabolic biomarkers as the most influential predictors, providing clinicians with practical, data-driven insights to support personalized recurrence risk assessment and targeted patient management. This work underscores the expanding role of machine learning in postoperative management, supporting personalized monitoring and targeted preventive strategies for patients undergoing ECIRS [7].

This study investigated the application of machine learning algorithms to predict symptomatic kidney stone formation using data from the Fasa Adults Cohort Study (FACS). Several models, including decision tree and ensemble-based techniques, were assessed based on demographic, clinical, and lifestyle attributes to identify individuals at a higher risk of stone development. The findings revealed a strong predictive performance, highlighting metabolic and lifestyle factors as the most influential determinants of kidney stone occurrence. This study highlights the importance of leveraging population-based cohort data to develop preventive, non-invasive prediction tools, that support earlier detection and more effective management of patients at risk of nephrolithiasis [8].

This study examined the application of machine learning models to predict 24-hour urinary abnormalities linked to kidney stone disease, utilizing data extracted from electronic health records (EHRs). Through the evaluation of multiple algorithms, the authors demonstrated that ML techniques can effectively identify patients at risk for hypercalciuria, hyperoxaluria, and hypocitraturia without relying on traditional urine collection methods. Ensemble approaches yielded the highest overall predictive performance, whereas feature importance analysis highlighted the most influential demographic and clinical variables contributing to risk identification. This research highlights the potential of ML as a non-invasive, cost-effective alternative for metabolic risk assessment, improving diagnostic efficiency and supporting preventive strategies in nephrolithiasis management [9]

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

A deep learning framework was developed for automated detection and volumetric segmentation of kidney stones in non-contrast CT scans. Trained and validated on an extensive dataset, the model demonstrated high accuracy in identifying stones of various sizes and anatomical regions, while delivering precise volumetric measurements. By minimizing dependence on manual interpretation, this approach enhances efficiency, reproducibility, and scalability of clinical workflows. This research demonstrates the potential of deep learning as a robust computer-aided diagnostic tool, supporting radiologists in stone evaluation and promoting standardized assessment of stone burden in both clinical and research settings [10].

This study explored the implementation of a deep learning model to assist in the detection of kidney stones on CT scans. The model was trained to automatically identify stones of different sizes and anatomical locations, and its performance was evaluated by expert radiologists. The findings revealed that the AI-assisted method enhanced accuracy and sensitivity, particularly in detecting smaller stones that are frequently missed. The study concluded that incorporating such tools into routine practice could improve diagnostic consistency, reduce human error, and accelerate clinical decision-making for kidney stone evaluation [11].

It introduced a deep segmentation network capable of simultaneously segmenting the kidneys and detecting kidney stones in unenhanced abdominal CT images. The authors developed an end-to-end framework that leverages multi-scale feature extraction and advanced segmentation methods to enhance the accuracy of both organ delineation and stone identification. The experimental evaluation demonstrated a high performance in detecting stones of varying sizes while maintaining detailed kidney segmentation. This approach improves diagnostic efficiency and scalability by minimizing manual annotation and streamlining image analysis. This work highlights the promise of deep learning–based segmentation models as effective, automated tools for accurate and clinically relevant nephrolithiasis assessment [12].

This work proposed a deep learning–driven framework for the automated detection and volumetric segmentation of kidney stones using non-contrast CT images. Researchers have designed a convolutional neural network (CNN) architecture that demonstrated high sensitivity and accuracy in stone identification. In addition to detection, the model provided precise volumetric measurements, marking a substantial improvement over conventional size-based assessment method. Validation using extensive CT datasets confirmed the model's robust performance, indicating its potential to reduce the workload of radiologists and enhance diagnostic consistency in clinical practice. Overall, this research highlights the potential of AI-driven tools in automating kidney stone evaluation to support clinical decision-making [13].

This study presents a deep learning framework for the automated detection and volumetric segmentation of kidney stones in non-contrast CT scans. Utilizing a convolutional neural network (CNN) architecture, the model demonstrated superior accuracy in stone identification and precise volume estimation, surpassing conventional thresholding and manual assessment techniques. It exhibited consistent performance across varying stone sizes and anatomical locations, underscoring its robustness and generalizability for clinical applications. By

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

reducing the radiologist workload and limiting interpretation variability, this approach highlights the potential of deep learning as a reliable computer-aided diagnostic tool to enhance efficiency and promote standardized evaluation of stone burden in clinical and research environments [14].

This study explored the use of machine learning models to predict kidney stone recurrence by integrating 24-hour urinary chemistry data with variables extracted from electronic health records (EHRs). A range of algorithms, including logistic regression, random forests, and gradient boosting, were evaluated, with ensemble-based methods consistently achieving the highest predictive accuracy. The analysis revealed that urinary markers such as calcium, oxalate, and uric acid, together with demographic and clinical features, were key determinants of recurrence. These findings emphasize the value of ML-driven tools in enabling personalized risk stratification, thereby enhancing preventive strategies and improving the efficiency of follow-up care for patients with nephrolithiasis [15].

Explore the use of temporally validated machine learning models to predict outcomes of percutaneous nephrolithotomy (PCNL), drawing on data from the British Association of Urological Surgeons (BAUS) PCNL Audit. Using this large, multi-center dataset, the authors developed and evaluated several ML algorithms to estimate surgical success rates and the likelihood of complications. The incorporation of temporal validation allowed the models to reflect evolving clinical practices and to enhance their reliability in real-world application. The results demonstrated that ML could deliver accurate and clinically valuable predictions, offering a foundation for personalized surgical planning, patient counseling, and tailored postoperative management in complex kidney stone cases [16].

Introduction of an AI-based machine learning framework to predict renal stone recurrence following Endoscopic Combined Intrarenal Surgery (ECIRS). The model integrated 24-hour urine chemistry profiles with postoperative clinical features and achieved strong predictive accuracy in identifying patients at an elevated recurrence risk. Urinary parameters have emerged as critical predictors of interpretability and clinical relevance. This work highlights the promise of AI-assisted postoperative risk stratification, supporting personalized follow-up, targeted preventive measures, and optimized long-term management strategies for patients with nephrolithiasis [17].

3. Methodology

i. Dataset Description

The dataset combines diverse variables including demographics, clinical and laboratory profiles, imaging findings, medical history, lifestyle factors, and treatment outcomes to support reliable kidney stone prediction. The dataset included demographic, clinical, imaging, medical history, and lifestyle information. Demographics included age, sex, and BMI; clinical variables included blood (creatinine, urea, calcium, uric acid, phosphate and electrolytes) and urine parameters (pH, calcium, oxalate, citrate, uric acid, phosphate, volume, and specific gravity).

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

Imaging data included stone presence, size, density, and location, while medical history captured family incidence, recurrence, and comorbidities such as diabetes, hypertension, CKD, gout, and obesity. Lifestyle factors included diet, fluid intake, and medication use. The outcome measures included stone type, recurrence, and treatment response. Overall, the dataset integrated continuous, categorical, binary, and survival-type variables, thereby facilitating comprehensive model training and evaluation.

ii. Data Preprocessing

Patient data were curated by combining demographic, laboratory, imaging, history, and lifestyle variables, while removing duplicates and incomplete records. Laboratory values were standardized, and imaging parameters were aligned for size, density, and location consistency. Missing data were handled using median/mode imputation, continuous features were normalized, and categorical variables were encoded. The outliers were adjusted within clinically valid ranges, and the derived ratios were calculated to reflect the lithogenic risk. To correct the class imbalance, class weighting and oversampling were applied to the training set only. All preprocessing operations were executed in a controlled pipeline to maintain data integrity and reproducibility.

iii. Machine Learning Models

A suite of supervised machine learning algorithms was implemented to assess predictive accuracy, robustness, and clinical relevance. Logistic regression served as the baseline for its interpretability via odds ratios and feature coefficients. Decision trees capture non-linear relationships and provide rule-based transparency, while random forests improved generalization and mitigate overfitting through ensemble aggregation.

Advanced ensemble models such as Gradient Boosting Machines (GBM) and XG Boost were utilized to handle heterogeneous clinical data and achieve high predictive performance. Support Vector Machines (SVMs) with optimized kernels evaluated high-dimensional separability, and k-Nearest Neighbors (k-NN) acted as a distance-based comparator to assess patient similarity.

Furthermore, Artificial Neural Networks (ANNs) have been explored to capture complex non-linear interactions across biochemical, imaging, and lifestyle variables. For recurrence prediction, survival-based variants of tree and regression models were applied to model the time-to-event outcomes. All models underwent uniform preprocessing, hyperparameter tuning, and cross-validation to, ensure consistent and reproducible comparisons across predictive strength, data resilience, and clinical interpretability.

iv. Evaluation Metrics

Model evaluation employed a comprehensive set of classification and robustness metrics to capture predictive accuracy, clinical reliability, and generalizability. Accuracy reflected overall

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

correctness, whereas precision, recall, and F1-score assessed the model's balance in distinguishing stone formation from non-stone patients. Given the high clinical risk of false negatives, recall has received particular emphasis, complemented by specificity to correct identify non-stone cases.

The Area Under the ROC Curve (AUC-ROC) quantifies the model's discriminative capability, illustrating the sensitivity specificity trade-off across thresholds. For imbalanced outcomes such as recurrence, the Area Under the Precision Recall Curve (AUC-PR) offered deeper performance insight. Calibration measures, including the Brier score and calibration plots, were analyzed to ensure alignment between predicted probabilities and observed outcomes, reinforcing the model's clinical interpretability and reliability.

The model robustness was further examined using cross-validation performance variability, site-level generalization (leave-one-center-out validation), and temporal validation of the data. For recurrence prediction, survival models were evaluated using the concordance index (C-index) and the time-dependent AUC. Collectively, these metrics ensured a rigorous, multi-dimensional assessment of predictive performance, robustness across settings, and clinical applicability.

4. Mathematical Modelling of ML Techniques

i. Logistic Regression (LR)

In the Logistic Regression, the probability of a patient developing kidney stones was modeled as a logistic function of the input predictors. Given a feature vector $X = (x_1, x_2, \dots, x_n)$ the linear combination of predictors is:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_n x_n$$

The logistic (sigmoid) function transforms this linear score into a probability value between 0 and 1:

$$P(Y = 1|X) = \frac{1}{1+e^x}$$

ii. Decision Tree (DT) Model

A Decision Tree classifies patients by recursively splitting the dataset into subsets that are increasingly homogeneous with respect to the kidney stone status. At each internal node, the algorithm selects the feature and threshold that best separate the classes using an impurity measure.

Two common criteria are Gini Impurity and Entropy:

Gini Impurity:

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

$$Gini(S) = 1 - \sum_{1=1}^{c} (P_i^2)$$

Entropy:

$$H(S) = -\sum_{1=1}^{c} p_i \, Log_2(p_i)$$

where p_i is the proportion of samples in $class\ i$ within, $node\ s$ and c is the number of outcome classes

The optimal split for a feature A is determined by maximizing **Information Gain (IG)**:

$$IG(S,A) - H(S) - \sum_{v \in Value(A)} (P_i^2) \ \frac{|S_v|}{|S|} \ H(S_v)$$

iii. Random Forest (RF) Model

The Random Forest algorithm extends the concept of Decision Trees by incorporating an ensemble approach to enhance predictive accuracy and robustness in kidney stone prediction. Rather than depending on a single tree, it generates multiple decision trees using varied subsets of the training data and features, then aggregates their outcomes through majority voting for classification tasks or averaging for regression analyses.

Formally, given T trees $\{ h_1(X), h_2(X), \ldots . h_T(X)\}$ used in the binary classification.

$$y = mode\{h_t(X)\}_{t=1}^{T}\}$$

iv. Gradient Boosting Machine (GBM)

Gradient Boosting Machines (GBMs) have been employed to capture complex, non-linear patterns in kidney stone prediction by iteratively integrating multiple weak learners most often decision trees into a robust ensemble model. Unlike Random Forests, which build trees independently, GBM constructs trees in sequence, with each new tree correcting the errors of the previous ensemble by minimizing the loss function through a gradient descent.

Formally, the model is expressed as:

$$F_M(x) = \sum_{m=1}^{M} y_m h_m(X)$$

v. Support Vector Machine (SVM)

Support Vector Machines were applied to classify patients into stone and non-stone groups by constructing an optimal hyperplane that maximizes the margin between classes. Given the training data (x_i, y_i), where $x_i \in \mathbb{R}^n$ is the feature vector and $y_i \in \{-1, +1\}$ is the class label, the SVM optimization problem is:

$$\min_{w,b} \frac{1}{2} ||w||^2 \text{ s.t } \ y_i(w^T x_i + b) \geq 1, \forall i.$$

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

For non-separable cases, stack variables ξ_i are introduced, leading to a soft margin formulation.

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \text{ s.t } y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$$

The decision function is then defined as:

$$f(x) = \text{sign}\left(\sum_{i\in S} \alpha_i y_i K(x_i, x) + b\right)$$

Where S denotes the support vectors and $K(\cdot)$ represents the represents the kerel function.

vi. k-Nearest Neighbors (k-NN)

k-NN was included as a distance-based comparator to assess patient similarity in the feature space. Given a new $sample\ x$, the algorithm computes the distances $d(x, x_i)\ for$ all training samples and selects the $k\ nearest\ neighbors$. The predicted class was then assigned using majority voting:

$$\hat{y} = \arg \max_{c\in\{-1,+1\}} \sum_{i\in N_k(x)} 1(y_i = c)$$

where $N_k(x)$ denotes the set of $k$nearest neighbors of x

vii. Artificial Neural Networks (ANNs)

ANNs have been explored to capture complex non-linear patterns across biochemical, imaging, and lifestyle variables. An input $vector\ x \in \mathbb{R}^n$ is transformed through the hidden layers, where each neuron competes.

$$h_j = \sigma(\sum_{i=1}^{n} w_{ij}x_i + b_j)$$

with σ. as the activation function. The final output layer produces the predicted class:

$$\hat{y} = \text{softmax}\left(\sum_j v_j h_j + c\right)$$

Algorithm 1: Data Preprocessing Pipeline

Input: Raw dataset D with features X and label y

Output: Cleaned, encoded, and scaled feature set X label y

Step 1: Remove exact duplicate rows from dataset D

Step 2: Drop features whose missingness exceeds a predefined value we considered >40%

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

Let dataset D have feature matrix $X = [f_1, f_2, \ldots f_n]$ with n samples

For each feature f_j:

$$MissingRate(f_j) = \frac{\#\{x_{ij} | x_{ij}=0, i=1,\ldots n\}}{n}$$

fraction of entries in feature f_j that are missing

Here we defined τ value =40%

$F_{drop} = \{f_j | MissingRate(f_j) > \tau$

Step 3: Handle the missing values.

We used categorized features f_j€$\{c_1, c_2, \ldots . c_k\}$, we define the input values as

$$x_{ij} = \begin{cases} x_{ij,} & if\ x_{ij} \neq 0 \\ C^J, & if\ x_{Ij} = 0 \end{cases}$$

Step 4: Encode the categorical variables.

For **sex, urine color, comorbidity yes/no → One-Hot Encoding (OHE).**

$$z_{i,j,l} = 1\{x_{i,j} = l\}$$

For stone composition (if >10 rare types of sub variations) smoothed target encoding with cross validation is

$$\theta_j(l) = \frac{n_l.\ Y_l^- + \mu\alpha}{n_l + \alpha}$$

Step 5: All numeric features were scaled using the Robertsdale, defined as

$$x_{i,j}^1 = \frac{x_{i,j} - Median(X_j)}{IQR(X_j)}$$

Step 6: Features with near-zero variance are removed.

For each feature X_j the variance was computed as follows

$$Var(X_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - X_j^-)\text{^}2$$

Algorithm 2 Robustness Evaluation

Step 1. Data Splitting

Cross-validation folds are stratified to maintain class proportion is

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

$$\frac{n+(v_{rk})}{|v_{rk}|} \approx \frac{n+(D)}{n}$$

Step 2. Preprocessing

Remove predictors with minimal variability that provide little discriminatory information.:

$$\text{Var}\,(x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2 \le \tau$$

Step 3. Model Training & Calibration

Calibrate the predicted probabilities through Platt scaling to improve calibration:

$$\overline{P_i} = \frac{1}{1+e^{-(\alpha_s \hat{\imath}+b)}}$$

Step 4. Base Metrics

The discriminative performance of the model was assessed by calculating the area under the ROC curve (AUC).:

$$A \cup C = \frac{1}{|P||N|}\sum_{i\in P}\sum_{j\in N}[\overline{P_i} > \overline{P_j}]$$

Step 5. Stress Testing

Simulate measurement variability by adding noise to test model stability

$$x'_{ij} = x_{ij} + \alpha\sigma_j\varepsilon_{ij}, \qquad \varepsilon_{ij}\sim N(0,1)$$

Step 6. Aggregation & Robustness Index

Aggregate performance into a composite Robustness Index (RI)

$$RI = \sum_j w_j\overline{m_j}$$

The robust evaluation framework examined kidney stone prediction models using six structured steps. It begins with stratified cross-validation to ensure a uniform class distribution within all the folds. Data preprocessing follows, where missing values are imputed, features are standardized, and variables with minimal variance are eliminated to enhance the data quality. Subsequently, each model was trained with tuned hyperparameters and calculated to ensure that the predicted probabilities reflected true clinical risk. Baseline performance was assessed using evaluation metrics including AUC, sensitivity, specificity, and calibration error. The robustness of the model was also examined under stress conditions, including added noise, simulated missing data, label perturbations, class imbalance, subgroup fairness checks, and stability of feature importance. Finally, the results are consolidated across folds with confidence intervals, and a comprehensive Robustness Index is derived to enable comparative assessment of model reliability.

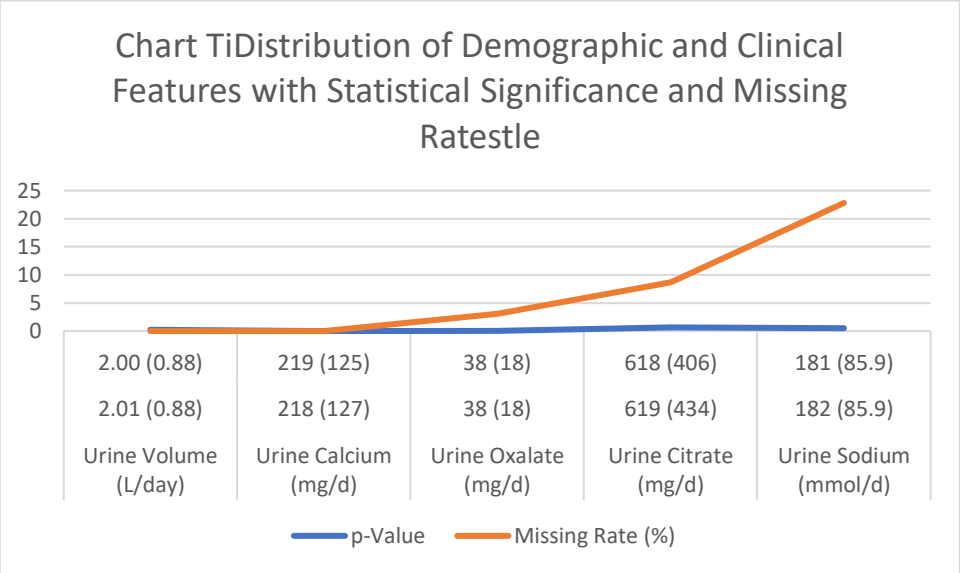
Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

5. Results and Discussion

The analysis of candidate features showed that the distribution of sex had no significant influence on stone formation, as both males and females were represented similarly in the stone and non-stone groups (p = 0.2363). However high blood pressure appeared to be an important factor, with a significantly higher prevalence in patients with stones (40.5%) than in those without stones (28.6%), supported by a p-value of 0.0158. In contrast, high cholesterol occurred more frequently among stone patients (57.4% vs 52.3%), but the difference did not reach statistical significance (p = 0.6563), and some data were missing (8.7%). Diabetes also showed a marginally higher occurrence in stone patients (10.1% vs 6.3%), however the difference was not significant (p = 0.5144), and the relatively high missing rate (22.8%) limited its reliability (Table 1).

Feature	No Stones (N)	Stones (Y)	p-Value	Missing Rate (%)
Gender (Female)	89	55	0.2363	0
Gender (Male)	198	93	0	0
HBP (Yes)	82	60	0.0158	3.1
High cholesterol (Yes)	150	85	0.6563	8.7
Diabetes (Yes)	18	15	0.5144	22.8

Table 1. Distribution of Demographic and Clinical Features with Statistical Significance and Missing Rates



Analysis of urinary biochemical parameters over 24 h showed that the values for patients with and without kidney stones were nearly identical. Urine volume averages 2.0 liters in both groups, indicating that overall fluid output did not differ significantly. Similarly, urinary calcium and oxalate levels which are key contributors to stone formation, remained almost the

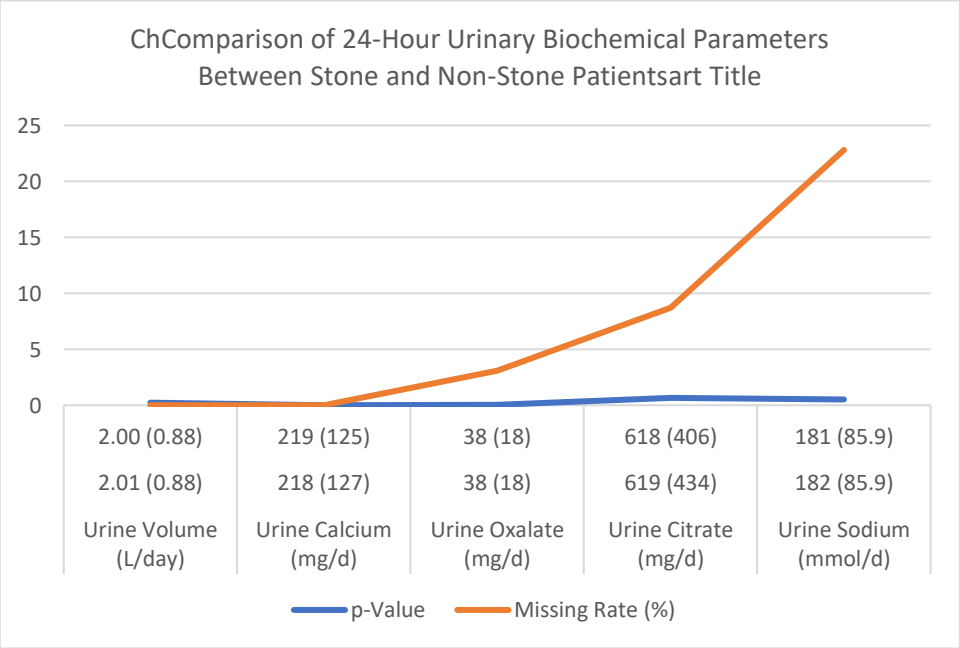
Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

same across the groups, suggesting that they may not distinguish stone risk in this dataset. Urine citrate, a known protective factor against stone formation, also showed no significant difference between groups (619 mg/day vs 618 mg/day). Likewise, urinary sodium levels were comparable, reflecting a similar dietary salt intake (Table 2).

Parameter	No Stones	Stones
Urine Volume (L/day)	2.01 (0.88)	2.00 (0.88)
Urine Calcium (mg/d)	218 (127)	219 (125)
Urine Oxalate (mg/d)	38 (18)	38 (18)
Urine Citrate (mg/d)	619 (434)	618 (406)
Urine Sodium (mmol/d)	182 (85.9)	181 (85.9)

Table 2. Comparison of 24-Hour Urinary Biochemical Parameters Between Stone and Non-Stone Patients

The comparison of the predictive models showed that all algorithms achieved moderate discrimination ability in identifying kidney stone risk. Logistic Regression yielded a validation AUC of 0.618, whereas LASSO performed slightly better at 0.625, suggesting that regularization improved predictive stability.



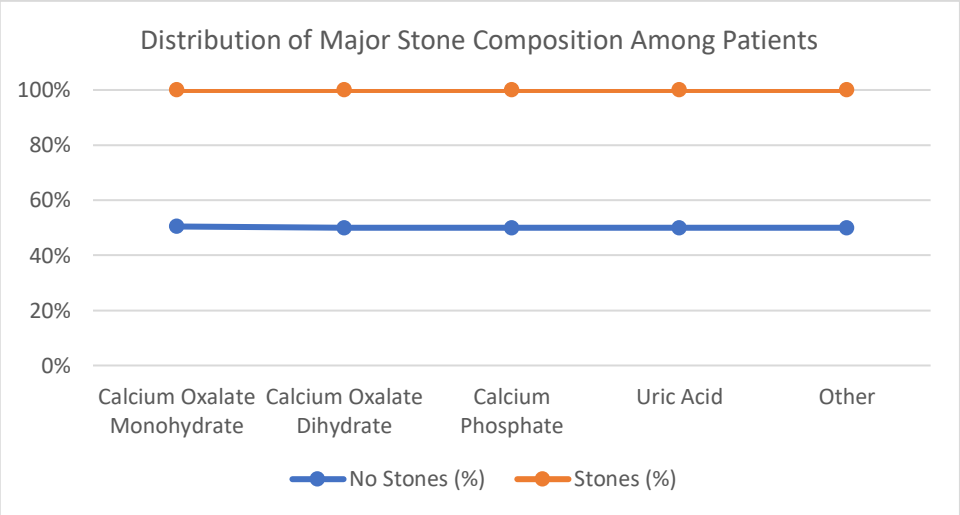
The distribution of stone composition among patients indicates that calcium-based stones are the most predominant type, with calcium oxalate monohydrate accounting for over half of the cases (55–56%), followed by calcium oxalate dihydrate (11%) and calcium phosphate (19%).

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

Uric acid stones constituted approximately 8% of the cohort, whereas the other types collectively represented approximately 7%. The proportions remained consistent across the groups, confirming that calcium oxalate stones were the most common form of nephrolithiasis, consistent with established clinical observations (Table 3).

Stone Type	No Stones (%)	Stones (%)
Calcium oxalate in its monohydrate crystalline form	56%	55%
Calcium oxalate in its dihydrate crystalline form	11%	11%
Calcium Phosphate	19%	19%
Uric Acid	8%	8%
Other	7%	7%

Table 3. Distribution of Major Stone Composition Among Patients



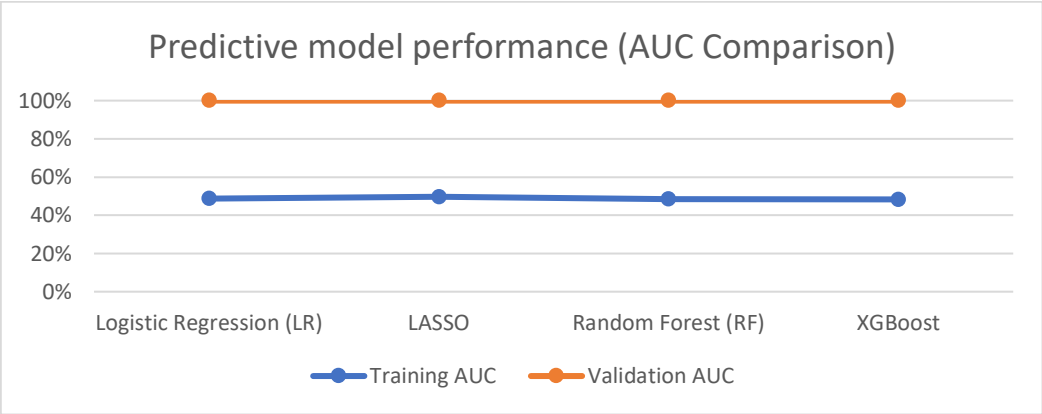
Random Forest achieved a validation AUC of 0.608, and XGBoost demonstrated similar performance at 0.621. Although the differences across models were relatively small, LASSO and XGBoost showed marginally better generalization than Logistic Regression and Random Forest. Overall, the results indicate that while machine learning methods provide some predictive value, their performance remains moderate, highlighting the need for enhanced feature engineering or the integration of additional clinical variables for stronger predictive accuracy (Table 4).

Model	Training AUC (95% CI)	Validation AUC (95% CI)
Logistic Regression (LR)	0.585	0.618

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

LASSO	0.617	0.625
Random Forest (RF)	0.570	0.608
XGBoost	0.580	0.621

Table 4: Predictive model performance (AUC Comparison)



6. Conclusion

This study demonstrated the potential of machine learning in predicting kidney stone disease, although the current models achieve only moderate accuracy. Hypertension emerged as a significant risk factor, whereas sex, cholesterol, diabetes, and standard 24-hour urinary biochemistry showed limited predictive value. Calcium oxalate stones were the most common type which is consistent with clinical observations. Among the tested models, LASSO and XGBoost offered slightly better generalization, whereas simpler models such as Logistic Regression provided higher interpretability for clinical use. Overall, the findings highlight the need for enhanced features and larger datasets to improve predictive performance, while emphasizing the balance between accuracy, robustness, and explainability for effective integration into clinical workflows

References

1. P. Doyle, W. Gong, R. Hsi, and N. Kavoussi, "Machine learning models to predict kidney stone recurrence using 24-hour urine testing and electronic health record-derived features," *Journal of Urology*, vol. 209, no. 3, pp. 657–664, 2023.
2. B. H. Eisner and D. S. Goldfarb, "A nomogram for the prediction of kidney stone recurrence," *Journal of the American Society of Nephrology*, vol. 25, no. 12, pp. 2878–2886, 2014.
3. R. S. Hsi, "Prediction tool to predict symptomatic kidney stone episodes: A step toward personalizing kidney stone care," *European Urology Focus*, vol. 4, no. 1, pp. 67–69, 2018.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

4. A. Abraham, N. L. Kavoussi, W. Sui, C. Bejan, J. A. Capra, and R. Hsi, "Machine learning prediction of kidney stone composition using electronic health record-derived features," *Journal of the American Society of Nephrology*, vol. 33, no. 8, pp. 1525–1535, 2022.
5. A. Abraham, N. L. Kavoussi, W. Sui, C. Bejan, J. A. Capra, and R. Hsi, "Machine learning prediction of kidney stone composition using electronic health record-derived features," *Journal of the American Society of Nephrology*, vol. 33, no. 8, pp. 1525–1535, 2022.
6. T. Yanase, R. Unno, T. Tokas, and V. Gauhar, "AI-driven prediction of renal stone recurrence following ECIRS: A machine learning approach to postoperative risk stratification incorporating 24-hour urine data," *Urolithiasis*, 2025.
7. T. Yanase, R. Unno, T. Tokas, and V. Gauhar, "AI-Driven Prediction of Renal Stone Recurrence Following ECIRS: A Machine Learning Approach to Postoperative Risk Stratification Incorporating 24-Hour Urine Data," *Journal of Clinical Medicine*, vol. 14, no. 12, article 4037, 2025.
8. F. Mahmoodi, A. Andishgar, E. Mahmoudi, A. Monsef, S. Bazmi, and R. Tabrizi, "Predicting symptomatic kidney stones using machine learning algorithms: Insights from the Fasa Adults Cohort Study (FACS)," *BMC Medical Informatics and Decision Making*, vol. 24, article 155, 2024.
9. N. L. Kavoussi, C. Floyd, A. Abraham, W. Sui, C. Bejan, J. A. Capra, and R. Hsi, "Machine learning models to predict 24-hour urinary abnormalities for kidney stone disease," *Journal of Urology*, vol. 208, no. 6, pp. 1218–1225, 2022.
10. D. C. Elton, E. B. Turkbey, P. J. Pickhardt, and R. M. Summers, "A deep learning system for automated kidney stone detection and volumetric segmentation on non-contrast CT scans," *Medical Physics*, vol. 49, no. 6, pp. 3664–3675, 2022.
11. A. Caglayan, M. O. Horsanali, K. Kocadurdu, E. Ismailoglu, and S. Guneyli, "Deep learning model-assisted detection of kidney stones on computed tomography," *Diagnostic and Interventional Radiology*, vol. 28, no. 3, pp. 213–218, 2022.
12. D. Li, C. Xiao, Y. Liu, Z. Chen, H. Hassan, L. Su, J. Liu, H. Li, W. Xie, W. Zhong, and B. Huang, "Deep segmentation networks for segmenting kidneys and detecting kidney stones in unenhanced abdominal CT images," *Frontiers in Medicine*, vol. 9, article 930144, 2022.
13. D. C. Elton, E. B. Turkbey, P. J. Pickhardt, and R. M. Summers, "A deep learning system for automated kidney stone detection and volumetric segmentation on non-contrast CT scans," *Medical Physics*, vol. 49, no. 6, pp. 3664–3675, 2022.
14. R. Manoranjitham, S. Punitha, V. Ravi, T. Stephan, A. Al Mazroa, P. Singh, M. Diwakar, and I. Gupta, "Automatic kidney stone detection system using guided bilateral feature detector for CT images," *Diagnostics*, vol. 14, no. 16, article 1757, 2024,

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 01-01-2026

15. P. Doyle, W. Gong, R. Hsi, and N. Kavoussi, "Machine learning models to predict kidney stone recurrence using 24-hour urine testing and electronic health record-derived features," *Journal of Urology*, vol. 209, no. 3, pp. 657–664, 2023,
16. R. M. Geraghty, A. Thakur, S. Howles, W. Finch, S. Fowler, A. Rogers, S. Sriprasad, D. Smith, A. Dickinson, Z. Gall, and B. K. Somani, "Use of temporally validated machine learning models to predict outcomes of percutaneous nephrolithotomy using data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit," *European Urology Focus*, vol. 8, no. 4, pp. 1001–1008, 2022,
17. T. Yanase, R. Unno, T. Tokas, and V. Gauhar, "AI-driven prediction of renal stone recurrence following ECIRS: A machine learning approach to postoperative risk stratification incorporating 24-hour urine data," *Journal of Clinical Medicine*, vol. 14, no. 12, p. 4037, 2025,