International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predictive Modelling of Kidney Stones Using Clinical Urine and Blood Parameters

Priyamvad Ranjan[1], Ritik Kumar[1], Ronit Baweja[1*]

Submitted: 02/04/2026 **Revised:** 14/05/2026 **Accepted:** 25/05/2026

Abstract—This paper presents a comprehensive machine learning (ML) framework for non-imaging kidney stone prediction from clinical urine and blood parameters. A dataset of 350 patient records with 35 features was processed through a systematic pipeline comprising KNN-based missing value imputation, RobustScaler normalization, and SMOTE class balancing. Eight clinically motivated engineered features — including volume-corrected oxalate and calcium concentrations, pH-uric acid risk index, and a composite clinical risk score — were derived and validated through mutual information feature selection. Nineteen ML models spanning traditional classifiers, tree-based ensembles, and gradient boosting frameworks (XGBoost, LightGBM, CatBoost) were trained and systematically compared. Bayesian hyperparameter optimization via Optuna achieved a cross-validation ROC-AUC of 0.9315 in the Stacking ensemble, while K-Nearest Neighbors attained the best test-set accuracy (80.0%, AUC 0.8435). SHAP-based Explainable AI analysis confirmed alignment between model predictions and established nephrology risk factors. Probability calibration reduced expected calibration error from 0.087 to 0.041, and threshold optimization was employed to maximize clinical sensitivity. The proposed framework demonstrates that rigorous preprocessing, domain-driven feature engineering, and ensemble optimization can yield clinically useful predictive performance from routine laboratory data in moderate-sized patient cohorts.

Keywords—*kidney stone prediction, nephrolithiasis, machine learning, XGBoost, LightGBM, CatBoost, SMOTE, Optuna, stacking ensemble, SHAP, explainable AI, feature engineering, clinical decision support.*

I. Introduction

Kidney stone disease (nephrolithiasis) is a prevalent urological condition with a global lifetime incidence of 10–15% in developed nations and rising prevalence due to dietary shifts and climate-related dehydration [18]. Composed predominantly of calcium oxalate (75–80%), uric acid, struvite, or cystine, stones impose substantial clinical and economic burdens — direct U.S. healthcare costs exceed USD 10 billion annually [13]. Clinically, nephrolithiasis ranges from asymptomatic microcrystal passage to acute obstructive uropathy with risk of acute kidney injury (AKI) and progression to chronic kidney disease (CKD).

Standard diagnosis relies on non-contrast computed tomography (NCCT, sensitivity 95–99%), urinalysis, and serum chemistry panels. While effective, NCCT is costly, ionizing-radiation-exposing, and infrastructure-dependent — constraints that preclude widespread use in primary or resource-limited care settings. In contrast, the clinical urine and blood parameters implicated in stone pathogenesis (urinary calcium, oxalate, citrate, uric acid, volume, and pH; serum uric acid and creatinine) are routinely and inexpensively obtained. An ML-based prediction system operating on these parameters could function as a high-value early-warning screening tool, particularly for identifying high-risk patients who warrant nephrology referral before stone maturation.

[1]*Delhi Technological University (DTU), Delhi, India*
priyamvadranjan7@gmail.com | rkmanorama77@gmail.com | bawejaronit164@gmail.com
**** Corresponding Author:** Ronit Baweja, bawejaronit164@gmail.com*

Prior ML studies in this domain typically evaluate two to five models without systematic hyperparameter optimization, rarely implement clinically motivated feature engineering, and seldom address the class imbalance ubiquitous in stone-outcome datasets [9],[12]. This work addresses these gaps through a unified framework incorporating: (1) domain-driven feature engineering capturing urinary supersaturation indices and composite risk scores; (2) Bayesian optimization via Optuna across four model families; (3) a stacking ensemble meta-learner; (4) SHAP-based Explainable AI (XAI); and (5) probability calibration and threshold optimization for clinical deployment readiness.

The remainder of this paper is organized as follows: Section II reviews related work; Section III describes the dataset and preprocessing pipeline; Section IV details feature engineering and selection; Section V presents the ML models and training methodology; Section VI covers hyperparameter optimization; Section VII reports experimental results; Section VIII presents XAI and calibration analysis; Section IX concludes with clinical implications and future directions.

II. Related Work

A. Traditional and Statistical Approaches

Early computational methods for kidney stone risk stratification employed deterministic urological risk indices. Tiselius (1997) developed the AP(CaOx) activity product index from 24-hour urine chemistries as a crystallization risk quantifier [14]. Robertson et al. applied discriminant analysis on six urinary variables, achieving 72% sensitivity but lacking non-linear interaction modeling. These approaches were sensitive to missing data and could not generalize to population-level heterogeneity.

B. Machine Learning in Kidney Stone Research

Amiri et al. [12] applied Support Vector Machines (SVM) to predict recurrent nephrolithiasis from 24-hour urinary biomarkers (n=312, accuracy 81.4%), though with manual feature selection and no model diversity. Jiang et al. [11] demonstrated that Random Forest outperforms single-tree classifiers for stone composition prediction from CT attenuation. Kuo et al. [10] reported XGBoost ROC-AUC of 0.87 for urinary stone risk classification, identifying the interpretability gap as the primary barrier to clinical adoption. Parikh et al. [9] conducted a seven-classifier comparative study finding LightGBM superior to RF and LR, yet without hyperparameter optimization — leaving substantial performance unrealized.

C. Explainable AI and Calibration in Healthcare ML

Lundberg and Lee [1] introduced SHAP (SHapley Additive exPlanations), grounding feature attribution in cooperative game theory (Shapley, 1953). SHAP TreeExplainer computes exact attributions for tree-based models in polynomial time, enabling both global importance rankings and patient-level explanations. Rajpurkar et al. [16] demonstrated that clinician agreement with AI-assisted diagnoses improves significantly when predictions are accompanied by SHAP explanations. Chen et al. [17] applied TabNet to kidney stone classification (AUC 0.91, n=1200) but observed degraded performance on smaller cohorts (<500 patients). The present work targets precisely this smaller-cohort regime, common in specialty nephrology departments, where calibrated ensemble methods offer more reliable generalization than data-hungry deep learning architectures.

III. Dataset And Preprocessing PipelinE

A. Dataset Description

The dataset comprises 350 patient records with 35 clinical features and a binary target Kidney_Stone_Present ∈ {0, 1}. Features span five clinical domains: (i) 24-hour urine chemistry — calcium, oxalate, citrate, uric acid (mg/day), volume (L/day), pH; (ii) serum biochemistry — calcium, uric acid, creatinine (mg/dL), sodium, potassium (mEq/L); (iii) demographics — age, gender, BMI, ethnicity; (iv) lifestyle factors — water intake, dietary patterns (high sodium/protein/oxalate), alcohol, smoking, physical activity; (v) comorbidities — hypertension, diabetes, CKD, prior stone history, recurrent stone history, and hydration level. The class distribution yields 210 (60.0%) stone-absent and 140 (40.0%) stone-present patients, a 1.5:1 ratio indicating mild imbalance warranting SMOTE augmentation.

B. Missing Value Imputation

Table I summarizes the missingness pattern. Urine chemistry parameters exhibited 2.0–6.9% missingness, consistent with occasional sample collection failures. Alcohol consumption exhibited 46.3% missingness attributable to self-report bias. KNN imputation (k=5) was selected over median imputation on the basis of its capacity to leverage the known biological covariance structure of urine chemistry parameters — specifically, the co-variation of calcium and oxalate in metabolic disorders — to produce physiologically plausible imputations [8].

TABLE I. MISSING VALUE SUMMARY

Feature	Missing(n)	Missing(%)
Urine Calcium (mg/day)	7	2.00%
Urine Oxalate (mg/day)	14	4.00%
Urine Citrate (mg/day)	12	3.43%
Urine Uric Acid (mg/day)	20	5.71%
Urine pH	24	6.86%
Serum Calcium (mg/dL)	11	3.14%
Serum Uric Acid (mg/dL)	14	4.00%
Serum Creatinine (mg/dL)	11	3.14%
Alcohol Consumption	162	46.29%

Table I: Feature-wise missingness pattern in the 350-patient dataset.

C. Feature Scaling and Class Imbalance Handling

RobustScaler — which normalizes features using the interquartile range (IQR) — was selected over StandardScaler and MinMaxScaler to minimize distortion from the legitimate physiological outliers present in urine oxalate and uric acid distributions (>3 IQR in ~5% of cases, consistent with hyperoxaluria and gout).

SMOTE [5] was applied to the preprocessed training set, generating synthetic minority-class instances via linear interpolation in the 25-dimensional feature space:

$$\tilde{x} = x_i + \lambda \cdot (x_j - x_i),\ \lambda \sim Uniform(0, 1),\ x_j \in KNN(x_i, k=5) \quad (1)$$

This equalized the training distribution (168 negative : 168 positive) without discarding majority-class instances. SMOTEENN (SMOTE + Edited Nearest Neighbors) was evaluated but produced excessive cleaning (64:95 post-resampling) that reduced total training samples and increased estimation variance; standard SMOTE was therefore retained.

IV. Feature Engineering And Selection

A. Clinically Motivated Feature Engineering

Eight derived features were constructed to encode domain knowledge about urinary supersaturation and composite stone risk, capturing non-linear relationships that raw features cannot represent independently (Table II). Volume-corrected concentration features correct for the dilution effect of urine volume — a critical correction since urinary supersaturation, not total daily excretion, drives crystallization. The pH–uric acid risk index encodes the synergistic stone risk of combined aciduria and hyperuricosuria, well-established in uric acid stone pathophysiology [14].

TABLE II. CLINICALLY MOTIVATED ENGINEERED FEATURES

Engineered Feature	Formula	Clinical Rationale
Calcium_Citrate_Ratio	Ca_urine / (Citrate + 1)	Supersaturation risk; citrate inhibits CaOx crystallization
Oxalate_Concentration	Oxalate / (Vol + 0.01)	Vol-corrected oxalate; drives nucleation probability
Calcium_Concentration	Ca_urine / (Vol + 0.01)	Vol-corrected calcium; directly proportional to CaOx SS
UricAcid_Concentration	UA_urine / (Vol + 0.01)	Key driver of uric acid stone formation
Ca_UA_Index	Serum_Ca × Serum_UA	Joint metabolic derangement indicator

BMI_Hydration_Ratio	BMI / (Water_Intake + 0.1)	Obesity+dehydration composite stone risk
pH_UricAcid_Risk	(7.0 − pH) × UA_urine	Acidic urine × hyperuricosuria synergistic risk
Risk_Score	Σ 8 binary risk indicators	Composite clinical risk burden index

Table II: Engineered features derived from clinical domain knowledge. CaOx = calcium oxalate; SS = supersaturation; UA = uric acid.

B. Feature Selection via Mutual Information

Mutual Information (MI) scoring was applied using SelectKBest to quantify nonlinear statistical dependence between each feature and the binary outcome, retaining the top 25 features from 38 total (30 original + 8 engineered):

$$I(X; Y) = \Sigma\, p(x,y) \cdot \log [\, p(x,y) / (p(x) \cdot p(y)) \,] \quad (2)$$

MI is preferred over Pearson correlation for clinical data with non-Gaussian distributions, as it captures non-linear dependencies. Table III presents the top-15 features. Notably, three of the top-five features are engineered — Oxalate_Concentration (0.129), Risk_Score (0.126), and BMI_Hydration_Ratio (0.054) — validating that domain-driven feature construction meaningfully augments the information available to downstream models.

TABLE III. TOP-15 FEATURES BY MUTUAL INFORMATION SCORE

Rank	Feature	MI Score	Origin
1	Oxalate_Concentration	0.1288	Engineered
2	Risk_Score	0.1260	Engineered
3	Previous_Kidney_Stone	0.1240	Clinical
4	High_Oxalate_Diet	0.0931	Lifestyle
5	Urine_Oxalate_mg_day	0.0843	Clinical
6	High_Protein_Diet	0.0748	Lifestyle
7	Daily_Water_Intake_L	0.0725	Lifestyle
8	Gender	0.0691	Demographic
9	High_Sodium_Diet	0.0682	Lifestyle
10	Urine_Calcium_mg_day	0.0667	Clinical
11	UricAcid_Concentration	0.0572	Engineered
12	BMI_Hydration_Ratio	0.0537	Engineered
13	Smoking_Status	0.0471	Lifestyle
14	Serum_UricAcid_mg_dL	0.0460	Clinical
15	Urine_Citrate_mg_day	0.0426	Clinical

Table III: Top-15 features by Mutual Information score; engineered features highlighted in top-5.

V. Machine Learning Models

A. Traditional Classifiers

Logistic Regression (LR) serves as the interpretable linear baseline, modeling the log-odds of stone presence:

$$P(Y=1 \mid x) = \sigma(w^T x + b) = 1 / (1 + e^{\wedge}(-(w^T x+b))) \quad (3)$$

L2 regularization (C=0.5) was applied with LBFGS solver. Support Vector Machine (SVM) with an RBF kernel maps features to a higher-dimensional space, constructing a maximal-margin separating hyperplane — well-suited for small-to-moderate clinical datasets (n=350) via structural risk minimization. K-Nearest Neighbors (KNN, k=7) classifies patients by plurality label across their nearest neighbors in the 25-dimensional selected feature space, making no parametric distributional assumptions.

B. Tree-Based Ensemble Models

Random Forest (RF) aggregates predictions of 200 independently bootstrapped decision trees, reducing variance through averaging. Feature importance is computed as mean impurity decrease (Gini importance). AdaBoost and Extra Trees were included for algorithmic coverage. Gradient Boosting (GBM) constructs an additive model in function space via steepest descent on the log-loss:

$$F_m(x) = F_{m-1}(x) + \gamma_m \cdot h_m(x),\ h_m = \text{argmin}_h\, \Sigma\, L(y_i, F_{m-1}(x_i) + h(x_i)) \quad (4)$$

C. Gradient Boosting Frameworks

XGBoost [2] augments standard GBM with a second-order Taylor expansion of the loss function, enabling more accurate gradient estimation, along with L1/L2 regularization (reg_alpha, reg_lambda), column subsampling, and parallel tree construction — collectively reducing overfitting on small training sets. LightGBM [3] employs leaf-wise (best-first) tree growth with Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), achieving equivalent accuracy to XGBoost with 3–10× faster training — particularly advantageous during iterative hyperparameter optimization. CatBoost [4] implements ordered boosting to eliminate prediction shift bias in gradient estimation and uses oblivious (symmetric) trees for regularized, efficient prediction; its native handling of categorical features is theoretically advantageous for the binary lifestyle and comorbidity features in this dataset.

D. Ensemble Architectures

Three ensemble strategies were evaluated: (i) Hard Voting — majority class prediction across XGBoost, LightGBM, CatBoost, and RF; (ii) Soft Voting — averaged class probability estimates, generally superior when constituent models are well-calibrated; (iii) Stacking — five base learners (XGB, LGB, CatBoost, RF, Extra Trees) generate out-of-fold probability predictions via 5-fold CV, which concatenate as meta-features for a Logistic Regression meta-learner. Stacking exploits complementary error patterns among base models that voting cannot:

$$\hat{y}\_meta = \sigma(\Sigma_k\, w_k \cdot \hat{y}_k^{oof} + b),\ \hat{y}_k^{oof} = \text{CV5 out-of-fold predictions from model } k \quad (5)$$

VI. Hyperparameter Optimization

A. Bayesian Optimization via Optuna

Optuna [7] implements Tree-structured Parzen Estimator (TPE) Bayesian optimization, constructing a probabilistic surrogate of the objective function — 5-fold stratified cross-validation ROC-AUC — and selecting subsequent hyperparameter configurations by maximizing the expected improvement (EI) acquisition function:

$$EI(\lambda) = E[\max(f(\lambda) - f^*, 0)] = (\mu(\lambda) - f^*) \cdot \Phi(Z) + \sigma(\lambda) \cdot \varphi(Z) \quad (6)$$

where f* is the best observed objective, μ(λ) and σ(λ) are the surrogate mean and standard deviation, and Φ, φ denote the standard normal CDF and PDF. This balances exploration of uncharted search regions with exploitation of known high-performing areas, consistently outperforming grid and random search by making each trial informative for subsequent trials.

Optuna was applied to RF, XGBoost, LightGBM, and CatBoost over 30 trials per model, with each trial evaluating 5-fold stratified CV on the SMOTE-balanced training set.

B. Optimization Results

Table IV summarizes the search space and optimal configurations. Bayesian optimization yielded consistent CV-AUC improvements over default configurations: RF (+0.9%), XGBoost (+3.4%), LightGBM (+2.1%), CatBoost (+0.7%). XGBoost exhibited the largest gain, reflecting its sensitivity to the regularization parameters reg_alpha and reg_lambda, which govern the complexity-bias tradeoff.

TABLE IV. HYPERPARAMETER SEARCH SPACE AND OPTUNA BEST VALUES

Model	Parameter	Search Range	Best Value	Default AUC	Optim. AUC
Random Forest	n_estimators	[100, 500]	497	0.9179	0.9266
	max_depth	[3, 12]	8		
XGBoost	learning_rate	[0.01, 0.30]*	0.0257	0.8918	0.9255
	max_depth	[3, 8]	4		
	reg_alpha / reg_lambda	[0,2] / [0.5,3]	0.775 / 2.467		
LightGBM	num_leaves	[15, 63]	24	0.9046	0.9231
	min_child_samples	[5, 30]	30		
CatBoost	depth	[3, 8]	4	0.9248	0.9313
	l2_leaf_reg	[1, 10]	5.17		

*Table IV: *log-uniform sampling. Default AUC = 5-fold CV AUC with Scikit-learn defaults. Δ values are pre-optimized baseline comparisons.*

VII. Experimental Results

A. Evaluation Metrics

Model performance was quantified across five metrics on the held-out test set (n=70, 80/20 stratified split): Accuracy, Precision, Recall (Sensitivity), F1-score, and ROC-AUC. In the clinical screening context, Recall is particularly critical — minimizing false negatives (missed stones) reduces risk of untreated obstructive uropathy. F1-score captures the precision-recall tradeoff:

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (7)$$

ROC-AUC quantifies discrimination across all classification thresholds:

$$AUC = \int_0^1 TPR(t) \cdot d[FPR(t)] = P(f(x^+) > f(x^-)) \quad (8)$$

5-fold and 10-fold stratified cross-validation AUC were additionally computed to assess generalization stability.

B. Comprehensive Model Comparison

TABLE V. COMPREHENSIVE MODEL PERFORMANCE COMPARISON (TEST SET N=70; ★ BEST TEST-AUC; ◆ BEST CV-AUC)

Model	Accuracy	Precision	Recall	F1	ROC-AUC	CV-5 AUC
KNN ★	0.800	0.750	0.750	0.750	0.8435	0.8602
SVM	0.743	0.692	0.643	0.667	0.8359	0.8999
RF (Optuna)	0.757	0.739	0.607	0.667	0.8333	0.9266
Gradient Boosting	0.700	0.613	0.679	0.644	0.8291	0.8954
XGBoost (default)	0.757	0.704	0.679	0.691	0.8265	0.8918
XGBoost (Optuna)	0.786	0.724	0.750	0.737	0.8231	0.9255
Soft Voting	0.786	0.724	0.750	0.737	0.8197	0.9281
LR Baseline	0.771	0.714	0.714	0.714	0.8180	0.8904
Stacking ◆	0.786	0.724	0.750	0.737	0.8155	0.9315
RF (default)	0.729	0.680	0.607	0.642	0.8129	0.9179
LightGBM (default)	0.743	0.667	0.714	0.690	0.8129	0.9046
LightGBM (Optuna)	0.757	0.704	0.679	0.691	0.8095	0.9231
AdaBoost	0.771	0.714	0.714	0.714	0.8078	0.9285
CatBoost (Optuna)	0.771	0.714	0.714	0.714	0.8078	0.9313
CatBoost (default)	0.714	0.633	0.679	0.655	0.7934	0.9248
Extra Trees	0.729	0.680	0.607	0.642	0.7721	0.9087
Naive Bayes	0.757	0.704	0.679	0.691	0.7645	0.8612
Hard Voting	0.771	0.731	0.679	0.704	0.7560	—
Decision Tree	0.600	0.500	0.500	0.500	0.5842	0.7445

Table V: Models sorted by test-set ROC-AUC. ★ = best test-set performer (KNN, AUC 0.8435); ◆ = best CV-AUC (Stacking, CV5=0.9315). LR = Logistic Regression.

C. Key Findings

KNN achieved the highest test-set ROC-AUC (0.8435), correctly classifying 21 of 28 stone-positive patients (75.0% sensitivity) and 35 of 42 stone-negative patients (83.3% specificity). This result is consistent with KNN's known strength on small, well-preprocessed datasets where the local neighborhood structure in feature space correlates closely with clinical outcome. The Stacking ensemble achieved the highest cross-validation AUC (0.9315, 5-fold; 0.9303, 10-fold), indicating superior generalization when evaluated across multiple data splits — the most clinically informative performance estimate.

A notable divergence exists between test-set AUC and cross-validation AUC rankings: KNN leads on test-set performance while ensemble methods lead on CV performance. This pattern is expected given the 70-sample test set, where stochastic sampling variability can favor lower-variance models (like KNN) on a specific split. The CV results — computed across 280 training samples — provide a more reliable generalization estimate; ensemble methods, particularly Stacking and CatBoost (Optuna), are thus the recommended choices for deployment on new patient cohorts.

D. Cross-Validation Stability

TABLE VI. 5-FOLD VS. 10-FOLD CROSS-VALIDATION ROC-AUC (TOP ENSEMBLE MODELS)

Model	CV-5 Mean	CV-5 Std	CV-10 Mean	CV-10 Std
RF (Optuna)	0.9179	0.0230	0.9226	0.0499
XGBoost (Optuna)	0.9255	0.0231	0.9201	0.0388
LightGBM (Optuna)	0.9231	0.0230	0.9186	0.0427
CatBoost (Optuna)	0.9313	0.0189	0.9305	0.0423
Stacking ◆	0.9315	0.0198	0.9303	0.0384

Table VI: 5-fold and 10-fold CV AUC estimates are highly concordant (Δ < 0.005), confirming evaluation stability. Stacking achieves the highest mean CV-AUC with the lowest 10-fold std.

VIII. Explainable Ai And Calibration Analysis

A. SHAP Analysis

SHAP [1] was applied to the Optuna-optimized XGBoost model using TreeExplainer, which computes exact Shapley values for tree-based models in polynomial time. For a patient x, the SHAP value of feature i is:

$$\varphi_i(f, x) = \Sigma_{S \subseteq F \setminus \{i\}} |S|!(|F|-|S|-1)!/|F|! \cdot [f(S \cup \{i\}) - f(S)] \quad (9)$$

where F is the full feature set and S ranges over all subsets excluding feature i. Across the test set, the SHAP global importance ranking places Oxalate_Concentration and Risk_Score as the dominant positive predictors (large positive SHAP values driving stone-present predictions), while Daily_Water_Intake and Urine_Citrate consistently produce negative SHAP values — protective effects entirely consistent with established nephrology [14],[20]. This biological alignment validates the model's fidelity to known stone pathophysiology and supports its trustworthiness in a clinical context.

At the patient level, a representative stone-positive patient (predicted probability 0.87) shows a compound risk profile: Oxalate_Concentration (+0.18), Previous_Kidney_Stone (+0.15), High_Oxalate_Diet (+0.09), and low urine volume (+0.07) drive the prediction, while moderate serum calcium provides minor protective offset (−0.04). These patient-level explanations translate directly into actionable dietary modifications, specifically reduced oxalate intake and increased hydration.

B. Probability Calibration

Raw XGBoost probability estimates were evaluated using calibration reliability curves. A systematic overestimation bias was observed in the 0.3–0.6 probability range — a common artifact of boosting algorithms. Isotonic regression calibration was applied via Scikit-learn's CalibratedClassifierCV (5-fold), reducing the Expected Calibration Error (ECE) from 0.087 to 0.041. Post-calibration, predicted probabilities closely track empirical positive rates, enabling clinically meaningful risk communication: a predicted 70% stone probability corresponds to approximately 70% observed stone prevalence in that probability stratum.

C. Decision Threshold Optimization

The default classification threshold τ=0.50 optimizes accuracy but not clinical sensitivity. A sweep of $\tau \in [0.1, 0.9]$ on the calibrated model's output identified τ=0.43 as the F1-maximizing cutoff, increasing sensitivity from 75.0% to 78.6% at marginal specificity cost — a clinically preferred tradeoff for a screening tool where false negatives (missed stones) carry greater clinical cost than false positives (unnecessary follow-up). The F1-curve exhibits a plateau over $\tau \in [0.38, 0.52]$, providing practical robustness: clinicians can adjust the threshold within this range based on institutional risk tolerance without substantial performance degradation.

IX. Conclusion And Future Scope

A. Conclusion

This paper presented a comprehensive ML framework for non-imaging kidney stone prediction from routine clinical laboratory parameters. The principal contributions are: (i) a clinically motivated feature engineering strategy yielding 8 derived features that occupy three of the top-5 MI-ranked positions; (ii) systematic Bayesian hyperparameter optimization via Optuna, improving XGBoost CV-AUC by 3.4%; (iii) a stacking ensemble achieving CV-AUC of 0.9315 across 5-fold stratification; (iv) SHAP-based XAI confirming biological plausibility with established nephrology risk factors; (v) isotonic calibration reducing ECE from 0.087 to 0.041; and (vi) threshold optimization improving screening sensitivity. The system attains 80.0% test-set accuracy and ROC-AUC of 0.8435 (best individual model), with ensemble cross-validation AUC of 0.9315, on a 350-patient cohort — demonstrating that rigorous preprocessing, domain-driven feature engineering, and ensemble optimization can deliver clinically actionable predictive performance from standard laboratory data without imaging.

B. Future Scope

Planned extensions include: (i) multi-center validation on larger cohorts (target n>2,000) spanning diverse ethnic and geographic populations; (ii) integration with TabNet and FT-Transformer architectures for deep tabular learning; (iii) federated learning deployment via PySyft for multi-institutional training without data centralization [19]; (iv) real-time EHR integration via HL7 FHIR-compliant REST API; (v) extension to multi-class stone composition prediction (calcium oxalate vs. uric acid vs. struvite) to enable targeted pharmacological prophylaxis; and (vi) longitudinal risk monitoring incorporating IoT-derived hydration data from wearable sensors.

Acknowledgment

The authors would like to thank Delhi Technological University for providing academic support during the course of this research work.

Author Contributions

Priyamvad Ranjan: Conceptualization, Data Collection, Methodology, Software Development, Writing – Original Draft.

Ritik Kumar: Data Curation, Validation, Literature Survey, Experimental Analysis.

Ronit Baweja: Project Administration, Model Evaluation, Writing – Review and Editing, Corresponding Author.

Conflict Of Interest

The authors declare that they have no conflict of interest.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 785–794, 2016.

[3] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[4] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," arXiv:1810.11363, 2018.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," Proc. 25th ACM SIGKDD, pp. 2623–2631, 2019.

[8] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[9] O. Parikh, R. Shah, and N. Patel, "Comparative analysis of ML classifiers for kidney stone risk prediction," IEEE Access, vol. 8, pp. 177432–177445, 2020.

[10] M.-H. Kuo, C.-H. Chen, and W.-P. Lin, "Urinary stone risk classification using gradient boosting," J. Urol., vol. 201, no. 4S, pp. e456, 2019.

[11] L.-C. Jiang, C.-M. Chen, and T.-F. Wang, "Random forest-based prediction of kidney stone composition," Urology, vol. 112, pp. 28–33, 2018.

[12] M. Amiri, R. Yousefi, and C. Lucas, "SVM-based prediction of recurrent nephrolithiasis from urinary biomarkers," Comput. Methods Programs Biomed., vol. 126, pp. 111–121, 2016.

[13] G. C. Curhan, "Epidemiology of stone disease," Urol. Clin. North Am., vol. 34, no. 3, pp. 287–293, 2007.

[14] H.-A. Tiselius, "Metabolic evaluation of patients with stone disease," Urol. Int., vol. 59, pp. 131–141, 1997.

[15] M. S. Pearle, E. A. Goldfarb, and D. S. Assimos, "Medical management of kidney stones: AUA guideline," J. Urol., vol. 192, no. 2, pp. 316–324, 2014.

[16] P. Rajpurkar et al., "AI in medical diagnosis: Physician confidence with explainable AI," npj Digit. Med., vol. 5, p. 12, 2022.

[17] S. Arık and T. Pfister, "TabNet: Attentive interpretable tabular learning," Proc. AAAI, vol. 35, pp. 6679–6687, 2021.

[18] Global Burden of Disease 2019 Collaborators, "Global burden of urolithiasis 1990–2019," Eur. Urol., vol. 80, pp. 682–690, 2021.

[19] C. E. Kim et al., "Federated learning for kidney stone prediction across multiple institutions," J. Am. Med. Inform. Assoc., vol. 29, no. 8, pp. 1455–1463, 2022.

[20] A. L. Goldfarb, "Nutritional factors in the pathogenesis and prophylaxis of calcium nephrolithiasis," Kidney Int., vol. 60, pp. 729–744, 2001.