AI-Driven Urine Analysis for Non-Invasive Kidney Stone Detection: A Machine Learning Approach

Praveen Kulkarni

Computer Science and Engineering
Dayananda Sagar University
Devarakaggalahalli , Ramanagar,
Bangaluru, 562112, Karnataka, India
praveen.kulkarni-cse@dsu.edu.in

Tejas M I

Computer Science and Engineering
Dayananda Sagar University
Devarakaggalahalli , Ramanagar,
Bangaluru, 562112, Karnataka, India
eng22cs0481@dsu.edu.in

Tejasvi D

Computer Science and Engineering
Dayananda Sagar University
Devarakaggalahalli , Ramanagar,
Bangaluru, 562112, Karnataka, India
eng22cs0482@dsu.edu.in

Venugopal

Computer Science and Engineering
Dayananda Sagar University
Devarakaggalahalli , Ramanagar,
Bangaluru, 562112, Karnataka, India
eng22cs0498@dsu.edu.in

Srujan S Shetty

Computer Science and Engineering
Dayananda Sagar University
Devarakaggalahalli , Ramanagar,
Bangaluru, 562112, Karnataka, India
eng22cs0469@dsu.edu.in

Abstract

The urological problem of kidney stones becomes a serious health concern when left untreated that may lead to unbearable pain and important medical complications. Patients often face two issues when using traditional diagnosis methods such as CT scans and ultrasounds: these tests cost a large amount of money and expose people to radiation while also being unavailable in areas that lack resources. The research evaluates a machine learning solution which examines urine test results to identify kidney stones because it aims to develop a cost-effective diagnostic method that operates through non-invasive methods with open accessibility.

Among the machine learning classifiers used in this study, Logistic Regression, Support Vector Machine (SVM), Random Forest and XGBoost were included because they are known to do well in biomedical classification. If your data is straightforward and linear, Logistic Regression is a simple method, but SVM deals well with more complex and unstructured data. Random Forest and XGBoost are types of ensemble models and they can manage both feature interactions and noise. While Random Forest helps more with smooth results and easy interpretation, it is XGBoost that stands out in making accurate predictions by applying gradient boosting.

Analysis of the urine was done using six important factors: specific gravity, pH, osmolality, conductivity, urea and calcium. The 1000 samples were checked for quality, normalized and looked into further to prepare for the model. Including interaction terms and clinical transformations into features improved how the model worked. Part of the data or 80%, was used for training and the remaining 20% was used for testing the classification model using accuracy, precision, recall, F1-score and AUC-ROC.

Random Forest had the best results, with an accuracy of 98.0%, precision of 0.989, recall of 0.968, F1-score of 0.978 and an AUC-ROC score of 0.986. XGBoost gave satisfactory results, but Random Forest did better at balancing sensitivity and specificity, making it the best option for early detection in the clinic. The results strengthen the idea that AI might help with kidney stone diagnosis without the need for imaging, especially when money is tight or proper care is far away.

By using machine learning with biochemical tests in a thoughtful manner, it is possible to improve detection earlier and to reduce the number of invasive procedures.

Keywords - non-invasive diagnosis, Random Forest, Logistic Regression, Support Vector Machine, XGBoost, urine analysis, kidney stone detection, machine learning techniques.

I. Introduction

The human kidneys create solid deposits of minerals and salts identified as kidney stones because of dehydration together with metabolic imbalances and dietary choices. Prolonged uncertainty about stones in the kidneys results in intense pain along with urinary tract infections which potentially cause significant harm to the kidneys unless quick medical treatment occurs. Traditional methods for kidney stone examination include CT scans together with ultrasounds and X-rays. These diagnostic techniques are expensive while additionally they subject patients to radiation exposure and also face limited availability especially in places with limited resources. The requirement exists for affordable and non-invasive approaches for diagnosis.

The current development of artificial intelligence enables image-based analysis to become an advanced tool that detects kidney stones. Medical imaging data analysis performs well for kidney stone detection and classification using Convolutional Neural Networks which combines machine learning and deep learning techniques on CT scans and ultrasounds data. Automated technology systems employ their programming to detect stone features as well as stone types which results in better diagnosis accuracy and less human interpretation dependence. AI-powered solutions provide a practical imaging methods replacement since they enhance the detection accuracy of kidney stones by using image processing methods that involve segmentation and feature extraction and image enhancement techniques.

Our study shows a new method for detecting kidney stones using only typical urine test results which means no CT scans or ultrasounds are needed. We add specific clinically relevant interaction terms (e.g., urea * calcium^{2}) and concentration ratios to our data to boost the model's ability to forecast outcomes and interpretability in practice. Logistic Regression, Support Vector Machine (SVM), Random Forest and XGBoost were trained and tested because they are known to work well in biomedical classification. edaw developed to investigate the distribution and relationships among features and performance was measured using standard metrics such as Accuracy, Precision, Recall, F1-score and AUC-ROC. By the results, it is clear that Random Forest does better than the others, so it can be trusted in real-life clinical use.

The research reveals machine learning tools demonstrate prospects to function as an economical alternative to knife diagnostics that enhances preliminary diagnosis outcomes alongside providing trustworthy alternatives to conventional testing approaches. Clinical staff will add further development to execute clinical validation tests before implementing the system at standard healthcare clinics.

II. RELATED WORK

Imaging techniques are the main way recent AI advancements are used to detect kidney stones. CT or ultrasound images were classified with high precision (over 90%) by researchers S. et al. [1], Manoranjitham et al. [2] and Nisha et al. [3], who employed CNN-based deep learning. Even so, they require excellent imaging machines and expose patients to radiation which means they are not ideal for places where resources are lacking. Unlike radiology approaches, we base our research on non-imaging urine chemistry which is simpler and safer.

Works like [2] and [4] apply machine learning models like SVM and DenseNet to the output of image processing methods. They are skilled in finding problems but they also depend on powerful imagery and advanced ways of analyzing features. We do not rely on CT images because our method uses freely available urine tests to make faster and more affordable diagnosis.

Some works such as Osei et al. [5] and Shee et al. [6], rely on urine tests and ML to identify kidney stones or to predict the recurrence of kidney stones. Even though they can be encouraging, these models tend to not use all the important features and do not compare many different models. Examples of this are seen in Osei et al. who only looked at osmolality and urea and in Shee et al., who achieved an AUC of 0.65. Including six key urine markers and their combinational terms gives us a larger feature set, resulting in improved diagnosis.

The work done by Alghamdi et al. [7] and Alqahtani et al. [10] introduced the use of Binary Particle Swarm Optimization (BPSO) with XGBoost on testing data. These approaches managed good results, though they spent most of their effort on tuning the algorithms and not exploring the relationships between features. Rather, we compare four different ML models, analyze the data deeply and apply clinical understanding to guide new features, so we can choose the best and most widely usable model.

In this way, our study adds to earlier work by establishing an extensive, interpretable and accessible ML process for catching kidney stones in urine tests.

III. METHODOLOGY

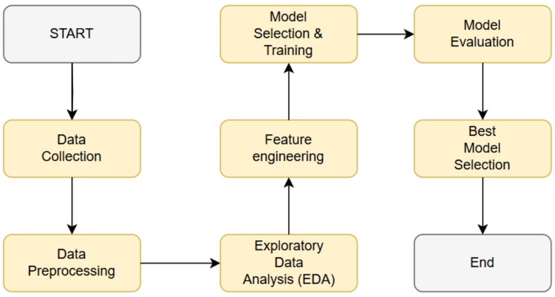
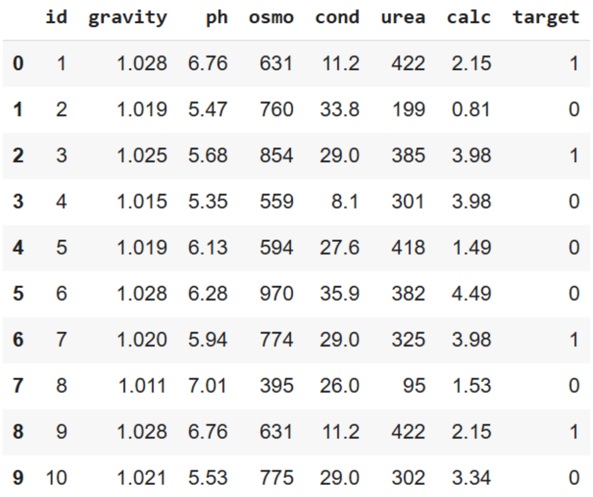


Figure 1. Work flow of the methodology

1. **Data Collection**: Collected urine test records which contained biochemical measurements derived from patient samples. The six essential physiological aspects that are measured in the analysis are specific gravity along with pH, osmolality, conductivity as well as urea and calcium. The dataset contains a two-value target field that shows whether a particular urinary condition exists or not. (Figure 2)

   In total, this investigation used one thousand anonymised pee test records. These are documents that came from a publicly accessible source chosen for research on kidney-related disorders. The dataset includes whether a patient has kidney stones and also records their specific gravity, pH, osmolality, conductivity, urea and calcium. Since the data was made safe for public release, it could be used without ethical clearance because no sensitive clinical info was included.



	id	gravity	ph	osmo	cond	urea	calc	target
0	1	1.028	6.76	631	11.2	422	2.15	1
1	2	1.019	5.47	760	33.8	199	0.81	0
2	3	1.025	5.68	854	29.0	385	3.98	1
3	4	1.015	5.35	559	8.1	301	3.98	0
4	5	1.019	6.13	594	27.6	418	1.49	0
5	6	1.028	6.28	970	35.9	382	4.49	0
6	7	1.020	5.94	774	29.0	325	3.98	1
7	8	1.011	7.01	395	26.0	95	1.53	0
8	9	1.028	6.76	631	11.2	422	2.15	1
9	10	1.021	5.53	775	29.0	302	3.34	0

Figure 2. Features and information about the dataset

2. **Data Preprocessing**: Preliminary processing techniques were extensively applied to the dataset to achieve data cleaning along with transformation, normalization and imbalanced data resolution as well as several other techniques because real-world data typically does not match ideal machine learning requirements. The input data contains noise together with missing data that needs processing before developing accurate predictions.

3. **Exploratory Data Analysis**: The dataset underwent a thorough exploratory data analysis for extracting vital information about its nature.

The figure shown in Figure 3 displays the target variable distribution where samples with (0) or (1) values represent kidney stone presence or absence respectively. The diagram displays information demonstrating 52.8% of the data points belong to class 0 and 47.2% belong to class 1. The proportional distribution between classes indicates that no significant changes need to be applied to the dataset for training effective machine learning models.

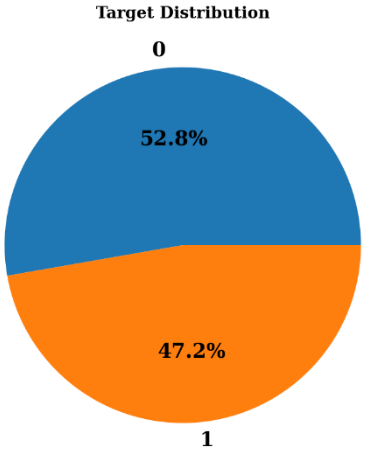


Figure 3. Stone distribution in dataset

The boxplot analysis in Figure 4 indicates that kidney stone patients demonstrate higher median values for all variables except pH including gravity, osmolality (osmo), urea and calcium concentrations (calc) in their urine samples. The difference between calcium as well as the urea_calc_interaction variable stands out as the most substantial indicating a potential link to stone development. The overall pattern in research data demonstrates that kidney stone occurrence appears linked to rising urinary concentration levels and distinct biochemical indicators although pH, conductivity (cond) and osmotic ratio measurements also vary mildly.

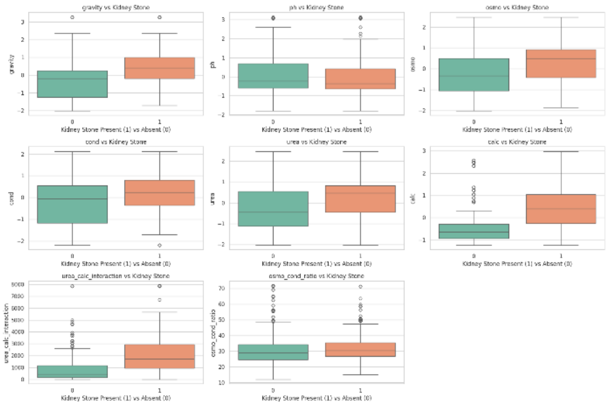


Figure 4. Boxplots of dataset feature (0-no stone and 1-stone)

The correlation heatmap (Figure 5) enabled us to determine how different urinary parameters relate to kidney stones. Data shows that urine calcium (calc) presents the highest positive association (r = 0.48) with kidney stone formation thus acting as a critical predictor. The risk factors for kidney stones appear to rise with higher concentrations of urine according to correlations with urea (r = 0.28) and specific gravity (r = 0.34). Urine pH demonstrates a minimal inverse connection (r = -0.12) so it contributes to stone formation to a lesser degree. Further evidence supporting the stone development process through concentrated urine came from the heatmap's display of strong connections between osmolality and conductivity as well as urea. The research results enable us to discover essential features which will be useful for developing exact machine learning model systems.

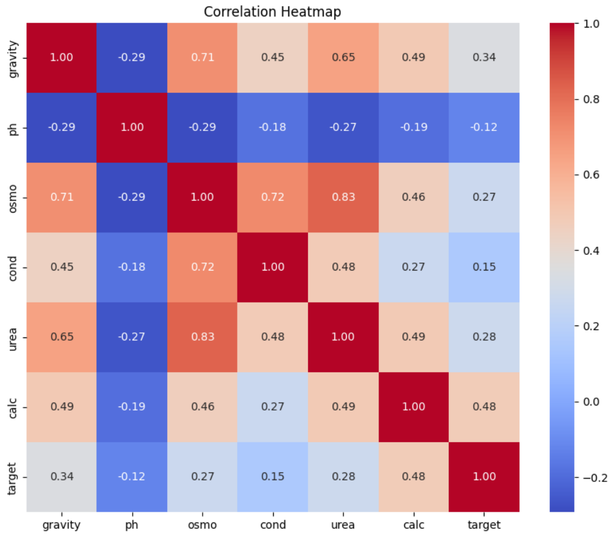


Figure 5. Heatmap correlation of dataset features

The pairplot analysis (Figure 6) indicates that urea, osmo and cond reveal the best capability for distinguishing target classes because class 1 displays separate clustering areas. The distributions of calc and gravity display average discrimination but ph data lacks effective separative capacity because both target classes overlap in the same areas. The correlation analysis shows osmo, cond and urea products tend to influence each other.

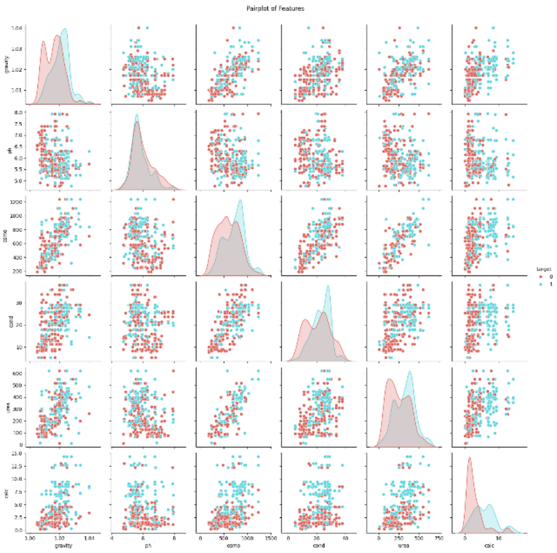


Figure 6. Pairplot of feature distribution

4. **Feature Engineering**: The model needed to better identify data patterns through implementation of multiple feature engineering approaches. The research incorporated interaction terms because it needed to analyze non-linear patterns by utilizing urea_calc_squared (urea × calcium^2), osmo_urea_interaction (osmolality × urea), and cond_calc_ratio (conductivity ÷ calcium). The model combinations formed from clear correlations and their medical usefulness during the previous exploratory stage analysis.

   Urine pH obtained three clinically relevant categories from ≤5.5 (acidic), 5.6 - 6.5 (neutral) to >6.5 (alkaline) which were converted to one-hot encoding for model processing. We developed two additional markers to measure urine concentration called total_solid_score and calc_osmo_ratio which calculated the mean osmolality value with specific gravity and conductivity measurements and the calcium amount compared to osmolality measurements respectively. Reasoning for non-linear relations required the addition of features containing calc_squared and urea_sqrt terms.

   The irrelevant identifiers were eliminated prior to splitting the data into training and testing parts for subsequent scaling work to avoid data leakage effects. StandardScaler transformed all continuous numerical features to achieve consistent contributions during model training operations.

5. **Model Training**:

5.1 Logistic Regression.

Logistic Regression functions as a linear binary classification model which applies the logistic (sigmoid) function to evaluate binary outcome probabilities. The boundary decision exists as:

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \qquad (1)$$

The model utilized linear solver alongside different values of C regularization strength and penalty type l1 or l2 for small dataset analysis. The model balanced class distribution through the use of class_weight = 'balanced'. Recall served as the primary metric for evaluation because the diagnostic system needed to detect all possible medical risks.

5.2 Support Vector Machine

With SVM we gain a hyperplane which creates the largest possible distance between two classes within a dimensional space. The optimization objective is:

$$\min \frac{1}{2}||w||^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1 \qquad (2)$$

SVM with rbf and poly kernels was used for detecting non-linear patterns while performing classification. Standardization ran as a pipeline to make all features equally impacting the result. A grid search method selected the regularization parameter C alongside the kernel coefficient γ (gamma). The introduction of balanced class weights served as a solution to deal with uneven distribution of samples.

5.3 Random Forest

Random Forest implements ensemble learning through the generation of numerous decision trees which combine results by using majority vote. The method serves to improve generalization abilities along with minimizing overfitting. The prediction is:

$$\hat{y} = \text{mode}\ (T_1(x), T_2(x), \ldots, T_n(x)) \qquad (3)$$

Where the prediction from decision tree number n can be represented by $T_n(x)$. A process of optimizing model parameters including n_estimators, max_depth, and min_samples_split was performed. The system applied Balanced class weighting for distributing data points among positive and negative classes to achieve fairness.

5.4 XGBoost

XGBoost operates as a high-speed gradient boosting system which constructs sequential tree structures to find the minimum value in a loss function with incorporated regularization components.

$$L = \sum_i l(y_i, \hat{y}_i^{(t)}) + \sum_i \Omega(f_k), \;\; \Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \quad (4)$$

Where:
The chosen criteria of l stands as the loss function (i.e. log loss for binary classification).
The value of Ω serves as a penalty which controls the complexity of the model.
T is the number of leaves
λ and γ control regularization strength

The model parameters max_depth and learning_rate together with subsample and colsample_bytree required optimization during the modeling process. The training process employed early stopping and evaluation metrics for the purpose of avoiding overfitting.

6. **Model Evaluation**: Testing and learning sections composed 20% and 80% respectively of the complete dataset for assessing all classifier models. Following metrics were used to evaluate the behavior of classification systems in the study.

- **Accuracy:**

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+\ FP+FN} \qquad (5)$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP+F\ P} \qquad (6)$$

- **Recall:**

$$\text{Recall} = \frac{TP}{TP+F\ N} \qquad (7)$$

- **F1 Score:**

$$\text{F1 Score} = 2 \times \frac{(Precision \times Recall)}{(Precision \times Recall)} \qquad (8)$$

- **ROC and AUC Score:**
  The receiver operator characteristic (ROC) curve shows the true positive rate (TPR) and false positive rate (FPR) at different threshold levels through its probability curve. A binary classifier performs best when depicted through ROC visualization. The area under the curve (AUC) determines a summary of the ROC curve. A model shows enhanced performance when it maintains a greater AUC value. A perfect classifier has an AUC value of 1. Within the same framework the random prediction yields 0.5 as its AUC measurement.

7. **Best Model Selection**: Model performance evaluation depends on key performance metrics which help establish effective medical diagnosis. Medical diagnosis requires Recall (Sensitivity) for detecting true positive cases together with precision for ensuring accurate positive predictions. Both F1-score and AUC-ROC serve to measure model reliability by balancing recall with precision and by determining class separation accuracy respectively. A confusion matrix serves as a detailed system that reports precise numbers of both correct and incorrect predictions. The best model selection relies on detection of missed diagnoses through higher recall values and better classification ability reflected by higher AUC-ROC scores and balanced

F1-score and fewer false negatives combined with medical safety and practical utility for clinical use based on speed and interpretation ease for doctors.

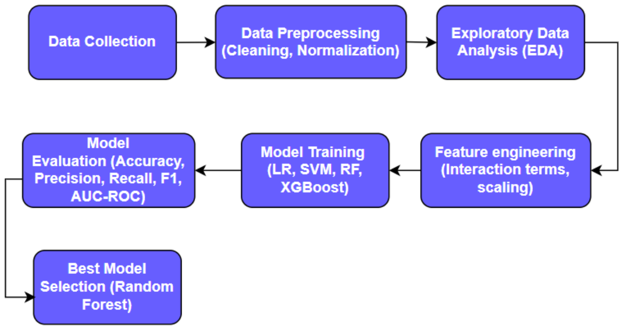

Data Collection
Data Preprocessing (Cleaning, Normalization)
Exploratory Data Analysis (EDA)
Feature engineering (Interaction terms, scaling)
Model Training (LR, SVM, RF, XGBoost)
Model Evaluation (Accuracy, Precision, Recall, F1, AUC-ROC)
Best Model Selection (Random Forest)

Figure 7. Proposed Kidney Stone Detection Pipeline.

Figure 7 illustrates the basic design of the suggested technic. Collecting urine test data is the start of the pipeline. Next, data undergo normalisation, get cleaned and the class distribution needs to be balanced. Following this, exploratory data analysis (EDA) is done to represent features in the manner they are related. Following this, interactions and ratios that are relevant for clinical purpose are found and produced by using domain-specific technology. A systematic way to divide a processed dataset into training and test parts is with an 80:20 ratio. Using grid search, these four machine learning models: XGBoost, Random Forest, SVM and Logistic Regression are trained and optimised. Finally, to decide on the best-performing classifier, the common performance metrics (Accuracy, Precision, Recall, F1-Score and AUC-ROC) are used to assess each model.

IV. Experimental Result And Discussion

Figure 8 presents confusion matrices for model classification that displays the important FP (False Positive) and FN (False Negative) metrics which are vital to medical diagnosis. Kennel stone detection outcomes with false positives forces patients through unnecessary treatment and anxiety but wrong negatives might delay correct diagnoses thus worsening patient results. The Random Forest (RF) classifier achieved the most outstanding performance metrics by maintaining only three false positive (3) and one false negative (1) results. The Logistic Regression (LR) model produced 28 false positives and 30 false negatives which stood as the highest numbers compared to other models and this indicates its lower reliability. The performance of Support Vector Machine (SVM) and XGBoost (XGB) fell between other models because SVM produced 15 FPs and 8 FNs while XGBoost generated 2 FPs and 3 FNs.

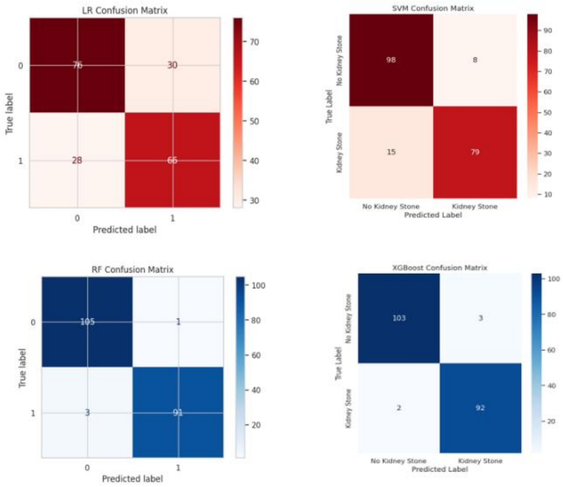
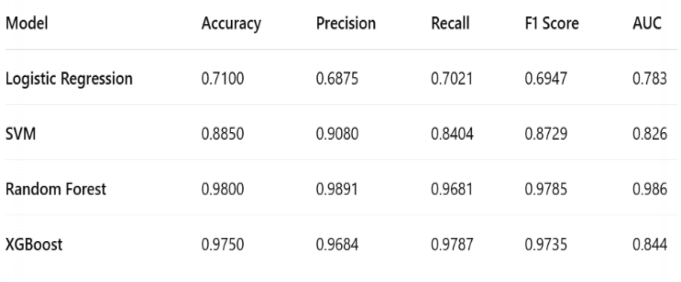


Figure 8. Confusion matrices of Models

Table 1 provides five important classification metric results including Accuracy along with Precision and Recall and F1-Score and AUC (Area Under the Curve). the formulas outlined in Equations (5) to (8) yield these numbers. The Random Forest model reached a performance score of 98% with Accuracy surpassing all other models including XGBoost at 97.50% and SVM at 88.50% as well as Logistic Regression at 71.00%. Random Forest achieved 0.96 Precision from its ability to identify and classify true positives from all positive predictions. The sensitivity of RF model exceeded 0.98 in its ability to detect positive cases during the Recall phase.

Table 1. Comparison of Models performance using valuation metrics



Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.7100	0.6875	0.7021	0.6947	0.783
SVM	0.8850	0.9080	0.8404	0.8729	0.826
Random Forest	0.9800	0.9891	0.9681	0.9785	0.986
XGBoost	0.9750	0.9684	0.9787	0.9735	0.844

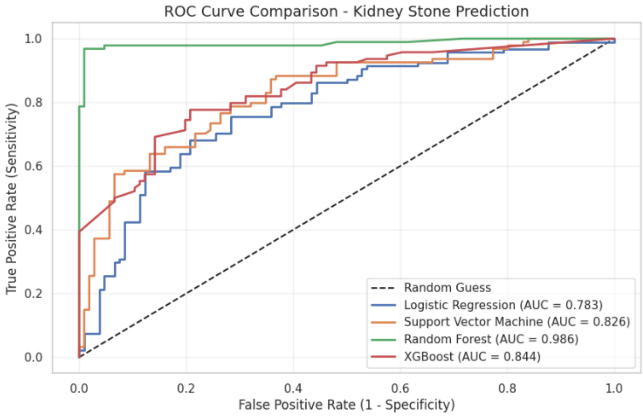


Figure 9. ROC Curve comparison of Models

The F1-Score achieved maximum value at 0.97 by RF which demonstrates that this model maintains optimal precision and recall values. The AUC value of XGBoost (0.844) compared to RF (0.986) does not translate into practical effectiveness measurement. The medical classification performance metrics demonstrate RF's strong suitability as the best solution for this task because it achieved superior results across all measures. SVM achieved 0.90 precision however its recall was at 0.84 which resulted in an F1-score of 0.87 while LR trailed behind with 0.68 precision and 0.70 recall and 0.69 F1-score.

When considering all evaluated metrics coupled with false negative reduction priorities the Random Forest approach excels as the most effective urine-based kidney stone detection solution.

Results from our model were compared with what was found in relevant studies to help understanding. To give an example, with a bigger feature set and specifically designed interactions, we achieved 98% accuracy, while Osei et al. [5] got a maximum accuracy of 94% using Random Forest analysis of urine parameters. While Shee et al. [6] tried to predict kidney stone recurrence with 24-hour urine data, their diagnostic capacity was regarded as poor, at 0.65 on the AUC. The AUC-ROC of Random Forest was also high, coming in at 0.986 which is better than the previous value. Alghamdi et al. [7] tried to use machine learning for identifying kidney stones in urine, though they did not assess and compare its different types, unlike our research. This method of comparing Logistic Regression, SVM, Random Forest and XGBoost allows our model to be scientifically precise and select the algorithm best fitted for clinical purposes. Since the model shows impressive recall, precision and is both simple and easy to use, it has good potential to be used for healthcare diagnoses in locations where imaging is either impractical or inaccessible.

V. Conclusion

Based on standard urine test measures, the study suggested a machine learning system for detecting kidney stones without needing invasive testing. When looking at accuracy, precision, recall and AUC-ROC, the Random Forest model proved to be the most accurate, reaching scores of 98%, 0.989, 0.968 and 0.986, respectively. I looked into and compared Logistic Regression, Support Vector Machine, Random Forest and XGBoost among many well-known classifiers. It became possible to get powerful and easy-to-understand results through both model evaluation and well-chosen medical features.

This research proves that using machine learning approaches can give reliable and inexpensive alternatives to standard imaging techniques when applied to organised biochemical analysis. The method was developed as an alternative when ultrasounds or CT scans are not accessible or can’t be used for financial reasons.

In coming studies, we want to prove that this method works with actual hospital data and measure its usefulness in practical settings. We also plan to implement complex methods such as deep learning and meta-learning and study developing a portable or web-based tool for the initial screening of health problems in places far from medical centers. All three aspects—clinical utility, interpretability and robustness—are expected to see important advancements because of these improvements.

References

[1] S, S., Ds, A., S, P., & Pai, V. (2023). Deep Learning Based Kidney Stone Detection Using CT Scan Images. 1–7.

[2] Manoranjitham, R., Punitha, S., Ravi, V., Stephan, T., Mazroa, A. A., Singh, P., Diwakar, M., & Gupta, I. (2024). Automatic Kidney Stone Detection System using Guided Bilateral Feature Detector for CT Images. *The Open Public Health Journal*, *17*(1).

[3] Nisha, N., Shrieya, S., Shreyas, R., Kulkarni, K., & Munavalli, J. R. (2024). Kidney Stone Detection using CNN Algorithm. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(6), 1814–1823.

[4] Priyadharsini, R., Balaji, V., & S., S. (2024). Deep Learning-Based Renal Stone Detection: A Comprehensive Study and Performance Analysis. *Applied Computer Systems*, *29*(1), 112–116. https://doi.org/10.2478/acss-2024-0014.

[5] Osei, I., Baafi-Adomako, A., & Boadu, D. (2024). Detecting Kidney Stones Using Urine Test Analysis: A Machine Learning Perspective. *International Journal of Research and Scientific Innovation*, *XI*(X), 754–771.

[6] Shee, K., Liu, A. W., Chan, C., Yang, H., Sui, W., Desai, M., Chi, T., & Stoller, M. L. (2024). A Novel Machine-Learning Algorithm to Predict Stone Recurrence with 24-Hour Urine Data. *Journal of Endourology*, *38*(8), 809–816.

[7] Alghamdi, H. S., & Amoudi, G. (2024). Using Machine Learning for Non-Invasive Detection of Kidney Stones Based on Laboratory Test Results: A Case Study from a Saudi Arabian Hospital. *Diagnostics*, *14*(13), 1343.

[8] Kavoussi NL, Floyd C, Abraham A, Sui W, Bejan C, Capra JA, Hsi R. Machine learning models to predict 24 hour urinary abnormalities for kidney stone disease. Urology. 2022 Nov 1;169:52-7.

[9] Degadwala, S., & Rathva, V. (2024). A Review on Kidney Stone Detection using ML and DL Techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(5), 101–112.

[10] Alqahtani, A., Alsubai, S., Binbusayyis, A., Sha, M., Gumaei, A., & Zhang, Y. (2023). Optimizing Kidney Stone Prediction through Urinary Analysis with Improved Binary Particle Swarm Optimization and eXtreme Gradient Boosting. *Mathematics*, *11*(7), 1717