Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260079
Published: June 9, 2026

Exploring Multivariate Techniques for Analyzing Kidney Stone Risk Factors: Logistic Regression and Cluster Analysis Approach

Omar Abueed
PhD Candidate
School of Systems Science and Industrial Engineering
Binghamton University (SUNY), USA
oabueed1@binghamton.edu

Audai Al-Akailah
PhD Candidate
School of Systems Science and Industrial Engineering
Binghamton University (SUNY), USA
aalakai1@binghamton.edu

Maya Abuarayes
Graduate Research Associate
School of Systems Science and Industrial Engineering
Binghamton University (SUNY), USA
mabuarayes@binghamton.edu

Mohammad Abueed, PhD
Module Development Engineer
Intel Corporation, USA
Mohammad.abueed@intel.com

Shane Menezes
PhD Candidate
School of Systems Science and Industrial Engineering
Binghamton University (SUNY), USA
smenezes@binghamton.edu

Yong Wang
Associate Professor
School of Systems Science and Industrial Engineering
Binghamton University (SUNY), USA
yongwang@binghamton.edu

Abstract

The aim of this endeavor is to explore the application of multivariate statistical methods in predicting whether the patient has kidney stone disease (KSD) or not. A public dataset has been obtained from Kaggle containing information on patients with KSD, including age, gender, and stone type, as well as laboratory test results. A systematic methodology has been followed, starting with exploratory data analysis (EDA) to identify and understand the patterns and trends in the dataset. As a second step, principal component analysis (PCA) has been applied. Then, based on the PCA results, a logistic regression (LR) model has been built. As a last step, clustering algorithms, including k-nearest neighbors and hierarchical clustering, were utilized to categorize the patients and run a comparison. The results showed that logistic regression was accurate in predicting the development of KSD with high precision, and the results of the clustering steps were two clusters. This research revealed the predictive potential of multivariate statistical methods in detecting KSD, and they have major consequences regarding the diagnosis and treatment of KSD patients.

Keywords

Machine Learning, Cluster Analysis, Kidney Stone Diagnosis, Logistic Regression, Principal Component Analysis (PCA).

1. Introduction

Kidney stone disease (KSD) is a urological disorder that can be characterized by hard crystalline deposits consisting of minerals and salts in the kidneys as a result of highly concentrated urine (Shastri et al. 2023). KSD in the abdomen is accompanied by intense pain that becomes worse during its passage through the urinary glands. The clinical presentation varies widely, ranging from asymptomatic cases diagnosed by several imaging procedures for other pathologies to acute renal colic characterized by sudden-onset severe flank pain, hematuria, nausea, and vomiting (Teichman 2004).

Globally, KSD prevalence ranges from 1% to 15%, with reported rates of 7%–13% in North America, 5%–9% in Europe, and 1%–5% in Asia, while underreporting remains common in low-resource regions (Stamatelou and Goldfarb, 2023). A recent epidemiological study found an increasing trend in the global prevalence of kidney stone incidence during the previous three decades, especially in developed nations, where the main key risk factors were dietary changes, obesity rates, and improved diagnostic capabilities (Stamatelou and Goldfarb, 2023). In the United States, KSD affects approximately one in every five hundred citizens annually, and the lifetime risk was estimated at 12.5% in men (peak incidence 40–60 years) and 6.25% in women (peak incidence 20–50 years). The economic consequences of KSD are substantial, and the annual healthcare costs were estimated at $10 billion US dollars. This cost comprises emergency department visits, surgical interventions, and long-term management of recurrent stone formers (Hyams and Matlaga, 2014).

The progression of KSD is related to different factors, including dietary practices, obesity measures, metabolic disorders, and urine anomalies. Kidney stone development is significantly exacerbated by the amount of animal salt and protein and oxalate consumption, with 2.5 times more sensitivity for those who have a positive heredity. Abnormal urinary parameters, including altered urine concentration, pH imbalance, and increased solute levels, enhance crystallization processes within the renal system (Borghi et al. 2002). This research aims to highlight clinical and laboratory elements that can be considered as associated risk factors with KSD prevalence, as well as to develop a predictive machine learning model using patient data.

2. Literature Review

In monitoring of the KSD and its associated risk factors, recent studies have been exposed to different types of methods, including traditional statistical methods such as the logistic regression method and advanced machine learning (ML) algorithms like the ensemble model techniques, with certain success rates under different clinical use cases and dataset features. The work of Abraham et al. (2022) investigated the effectiveness of XGBoost and LR models based on demographic and other medical characteristics coming from 1,296 patients to predict KSD composition. The work is considered a binary classification task, discriminating KSD patients with and without calcium. XGBoost achieved 91% accuracy compared to 71% for logistic regression, though both models reached similar ROC-AUC values of 0.80. For multiclass classification tasks, on the other hand, logistic regression was superior in terms of prediction accuracy compared with that of XGBoost; therefore, algorithm choice should depend on specific predictive outcomes.

An investigation conducted by Mahmoodi et al. (2024) analyzed data from 10,128 patients and deployed several ML algorithms, including Support Vector Machine, Random Forest, and XGBoost. The XGBoost model outperformed the other models and achieved an accuracy of 0.60, while the other algorithms achieved an accuracy ranging from 0.52 to 0.58. The dominant risk factors were determined to comprise the serum creatinine level, salt consumption, hospitalization history, and sleep pattern, highlighting the potential for early risk stratification through machine learning algorithms. Similarly, Iparraguirre-Villanueva et al. (2025) conducted an experiment to compare six ML algorithms for KSD detection and concluded that LR achieved the highest accuracy of 0.78. The LR outperformed other complex ML models, namely KNN (0.61), decision trees (0.66), SVM (0.68), Gaussian Naive Bayes (0.70), and random forests (0.76). This result implies that simpler models can be efficient for binary classification problems when feature relationships are predominantly linear. Doyle et al. (2023) examined the kidney stone recurrence prediction using Lasso, random forest, XGBoost, and LR models with 24-hour urine testing data from 1,231 first-time stone formers. The Lasso model had the best performance with AUC, with reported values of 0.617 and 0.625 for 2-year and 5-year recurrence predictions, respectively.

In a different study, in the southeast of Iran, the prevalence and major risk factors were investigated for KSD. The study included 9,932 participants aged 35-70 years, 25% of whom had a history of KSD. The study employed both univariable and multivariable LR analyses. The study determined the most dominant factors, such as gender, income, lack of purified water consumption, BMI, and history of hypertension and diabetes (Khalili et al. 2021). The theoretical foundation for combining principal component analysis with clustering was established by Husson et al. (2010), who proposed PCA as an effective denoising step leading to more stable clustering results. This approach is particularly helpful in high-dimensional medical datasets. The work of Ye et al. (2020) exploited the PCA procedure in clustering COVID-19 patients using K-means clustering approaches by ruling out 18 factors, rendering six principal components to explain 73.18% of variance. The resulting three clusters showed markedly different mortality rates, which demonstrated the clinical utility of unsupervised methods for patient stratification.

Another piece of research conducted by Maugeri et al. (2021) explored the introduction of hierarchical clustering with PCA for SARS-CoV-2 epidemiological pattern detection. They applied the PCA to these highly correlating variables and reduced them to three principal components that could capture 81.6% of total variance prior to running the hierarchical clustering. This approach combined PCA to realize the reduction of feature dimension with hierarchical clustering and K-means, which would get feasible clustering results in line with prior epidemic indexes. The silhouette coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index are among the cluster validation statistics. These criteria make it possible to have an additional measure to elect the best cluster number and assess cluster quality (Arbelaitz et al. 2013). Buch et al. (2024) proposed a scalable method that uses latent factor models with convex clustering penalties for biomedical data applications. Their hierarchically clustered PCA method outperformed fourteen traditional clustering approaches on datasets with high dimensionality and limited sample sizes.

Exploratory data analysis (PCA) remains essential for understanding distributional differences and feature correlations. Random Forest feature importance analysis complements traditional correlation analysis by capturing nonlinear relationships and interaction effects before the analysis begins. LR continues to be widely used in medical prediction due to its interpretability and reliability. The algorithm has coefficients that can be directly interpreted as log-odds ratios, facilitating translation of statistical findings into clinical recommendations, which ultimately ensure both separation ability and calibration across patient populations (Hosmer et al. 2013). The literature shows that combining several analytical methods, such as PCA for dimension reduction, logistic regression for interpretative prediction, and clustering, can assess patient subgroups. Thus, the complete framework is used for risk factors of KSD. The three approaches provide complementary values: PCA removes noise, explains variance, logistic regression can allow for statistical inference and clinical interpretability, and clustering identifies different phenotypes (patients) with varying biochemical characteristics.

3. Methodology

3.1 Dataset

The dataset is a binary classification problem and was obtained from a publicly available repository on Kaggle (https://www.kaggle.com/datasets/vuppalaadithyasairam/kidney-stone-prediction-based-on-urine-analysis). The data contains clinical records from 90 patients with seven variables, including a categorical dependent variable ("Risk of Stone") and continuous predictors such as urine specific gravity, urine pH, osmolality, urea concentration, and blood calcium levels. Prior to analysis, several steps were taken to validate the data. Completeness checks confirmed that no

missing values existed across any of the seven variables, and the class distribution was verified to be balanced (approximately 50/50 stone vs. non-stone), reducing the risk of sampling bias. Distributional plausibility was assessed by comparing the reported ranges of urine specific gravity (1.005–1.040), pH (4.76–7.94), osmolality (187–1236 mOsm/kg), and serum calcium (0.17–14.34 mmol/L) against established clinical reference intervals in the literature. Inter-variable correlations were also examined for physiological consistency; for example, the strong positive correlation between osmolality, conductivity, and urea is expected given their shared dependence on urinary solute concentration. While the original clinical source and collection methodology are not documented on the repository, the internal consistency of the data and alignment with known clinical ranges support its suitability for the exploratory multivariate analysis conducted in this study. Table 1 reports the KSD parameters and their associated definitions.

Table 1. Variables Definitions

Variable	Definition
Urine specific gravity (USG)	Measures the concentration of particles in urine compared with the density of water. Specific gravity results above 1.010 can indicate mild dehydration. The higher the number, the more dehydrated the patient is.
pH	Measures the concentration of uric acid in urine. The normal values range from pH 4.6 to 8.0. Above and below these ranges impact kidney functions.
Osmo	An osmolality test measures the concentration of particles in a urine sample. The unit of measure is mOsm/kg of water. Greater than normal results may indicate conditions such as Addison's disease, congestive heart failure, or shock.
Urine urea nitrogen (Urea)	Measures the amount of protein in the diet. Low levels usually indicate kidney problems or malnutrition (inadequate amount of protein in diet). High levels indicate an increase in the amount of protein breakdown in the body.
Blood calcium level (Calc)	Measures the calcium level. High levels of calcium weaken bones and form kidney stones.

3.2 Principal Component Analysis and Logistic Regression

PCA was applied to the six independent variables to reduce their total number and to find out which factor represents more variance of the data. Logistic regression analysis with the dependent variable of interest was conducted using the Fisher scoring optimization algorithm. The data was initially balanced for the two levels of the response variable.

3.3 Cluster Analysis

We applied two unsupervised clustering algorithms to reveal hidden patterns in the dataset. The data was preprocessed before clustering, where feature scaling and normalization were applied to have equal variances of the variables. In the end, k-means clustering and hierarchical clustering were used after systematic optimization. The optimal number of clusters was first calculated through the elbow method, evaluating the WCSS across different values of k. K-means models were assessed for cluster counts between two and ten, with the configuration at the inflection point of the WCSS curve identified as the optimum solution. The validity of the clusters was evaluated based on a visual inspection as well as quantitative measures of cluster performance. The silhouette coefficient was used to evaluate intra-cluster compactness and inter-cluster isolation, and a higher value indicates better clustering quality. Similarly, the Calinski–Harabasz Index was applied to measure cluster compactness and separation by computing the ratio of between-cluster deviation to within-cluster deviation as in Equation (1):

$$CH = \left(\frac{B}{k-1}\right)/\left(\frac{w}{n-k}\right)$$

Where B represents the sum of squares among clusters, W denotes the sum of squares within each cluster, k is the number of clusters, and n is the total number of observations. The Davies–Bouldin Index was utilized to measure the

average similarity between each cluster and its most similar neighboring cluster. Reduced Davies–Bouldin values signify enhanced clustering efficacy. The integration of these complementary evaluation metrics allowed a comprehensive and unbiased analysis of the clustering options.

4. Results and Discussion

4.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to assess the quality of data, distributions of features, statistical separability between result groups, and predictor importance before model fitting. The class distribution of the target variable was balanced; therefore, sampling bias in classification was reduced. Completeness checks confirmed that there was no missing data in any of the laboratory or clinical information, providing robustness for future modeling and statistical analysis. Figure 1 shows the dataset overview and distribution patterns of the six.

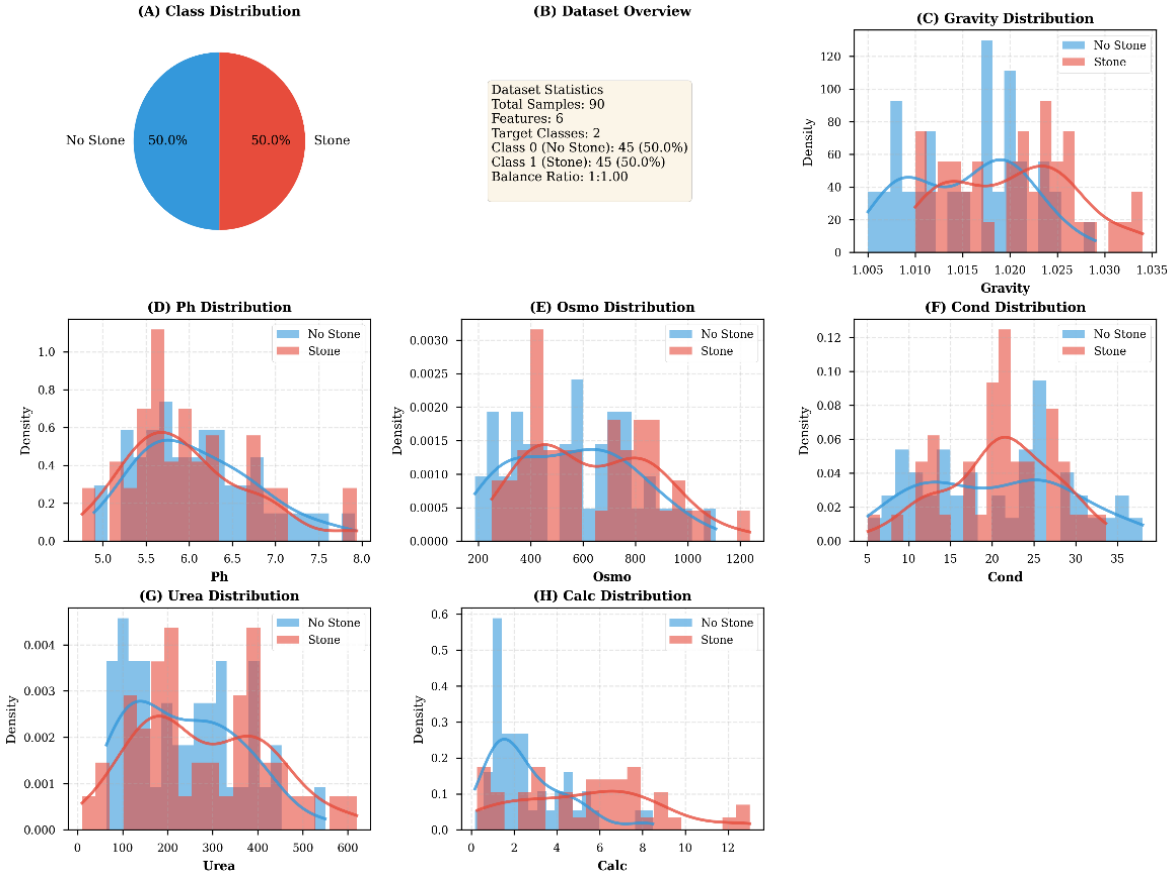

Figure 1. Comprehensive Exploratory Data Analysis of Kidney Stone Dataset

Statistical analysis by distribution using kernel density plots and boxplots demonstrates there are systematic differences between stone and non-stone groups. Median urine specific gravity and serum calcium were both significantly higher and more variably distributed in stone-positive patients, reflecting the known normal physiology of urinary supersaturation and crystal precipitation. Osmolality and urea demonstrated moderate alterations in distribution (skewed) and increased variability, which could reflect metabolic and hydration effects on stone development. There was more overlap for urine pH and conductivity between groups, which means a weaker univariate discriminatory power.

Strong relationships were observed by correlation analysis, in particular between osmolality, conductivity, and urea levels, which were to be expected given their common dependency on the concentration of solutes. Features-to-target correlations showed that calcium and urine-specific gravity had the highest positive association with the risk of stones, while pH had a weak inverse relation with it. Statistical hypothesis testing also confirmed that calcium and urine

specific gravity were significantly different among the groups ($p < 0.05$), while others contributed less individually to the discrimination among them.

To address non-linear relationships and interactions, which are not indicated in correlation measures alone, we also ran Random Forest feature importance analysis. Amongst the predictors that contributed to the discrimination model were calcium, urine SG, conductivity, and urea osmolality, thereby demonstrating the multivariate importance of these biomarkers. In general, the EDA results confirm the predictive capacity of these selected clinical features and support their integration with machine learning. Figure 2 presents the correlation and feature importance analysis. Figure 3 presents boxplot comparisons of urinary features between stone and non-stone patients.

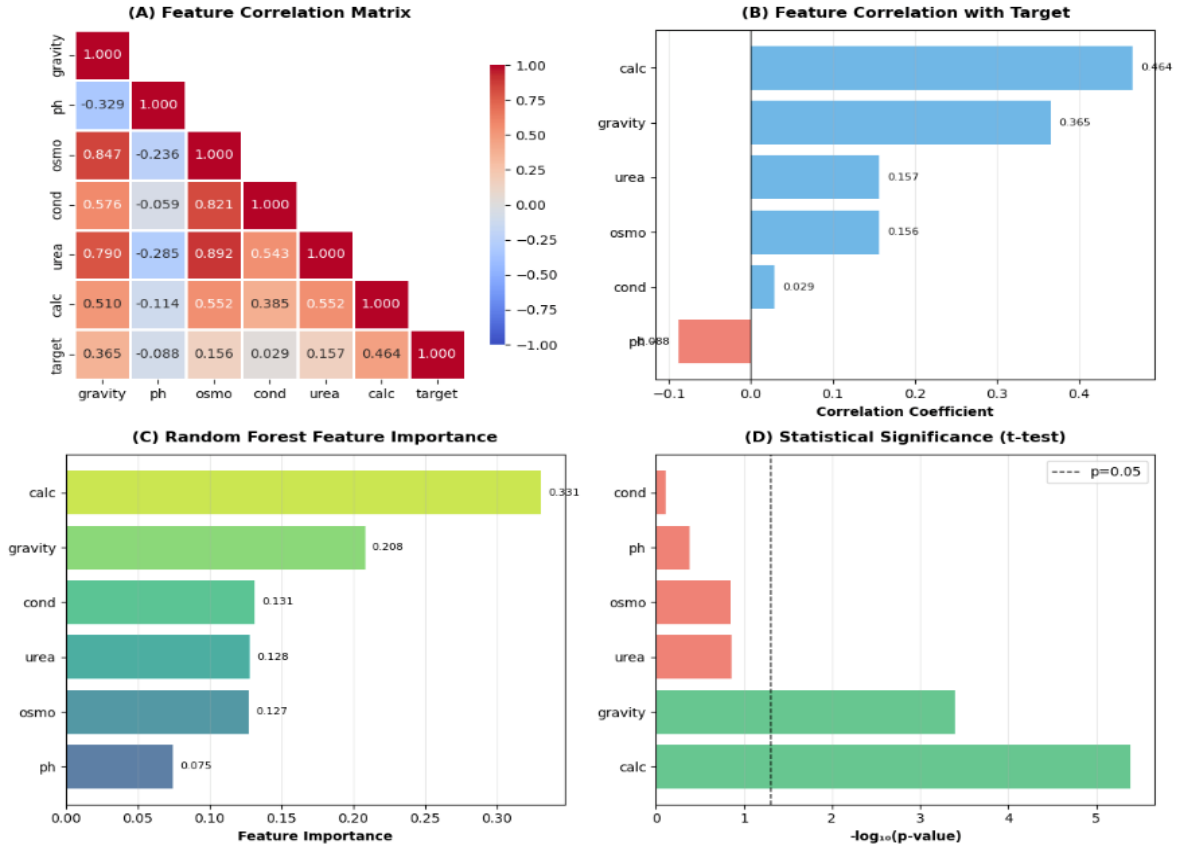

Figure 2. Feature Correlation and Statistical Analysis

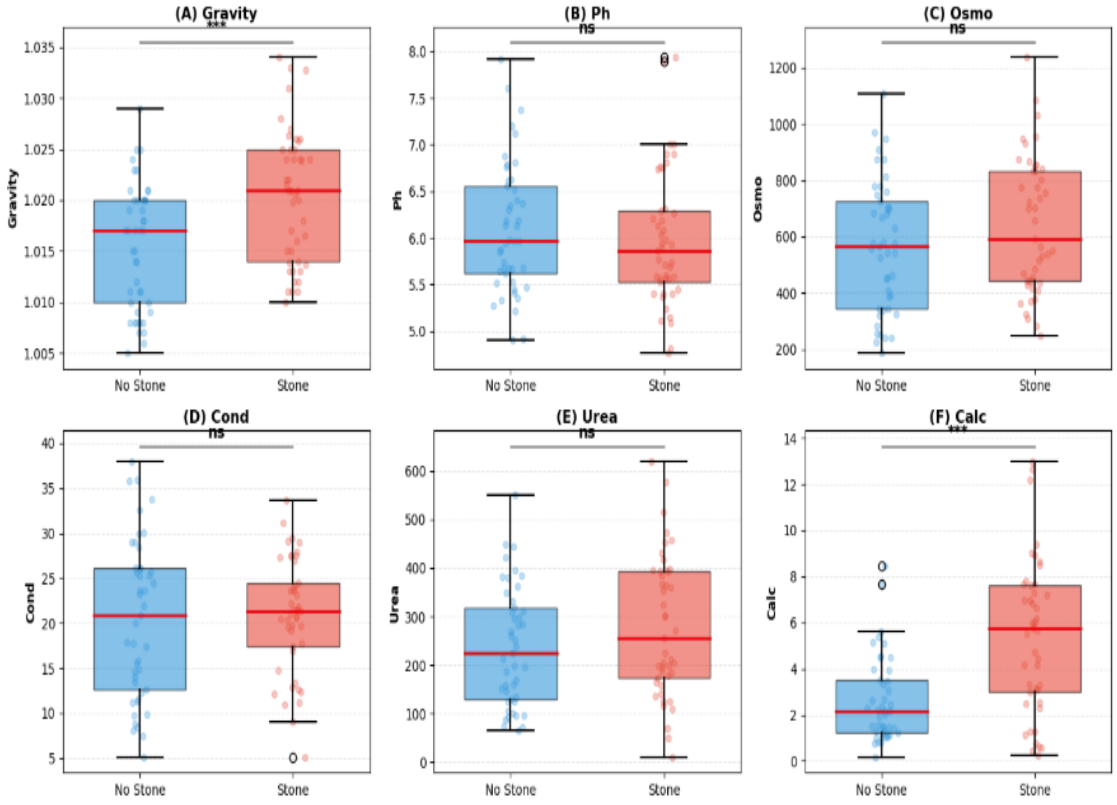

Figure 3. Feature Distribution by Class with Statistical Significance

4.2 Principal Component Analysis (PCA) and Logistic Regression

PCA has been applied to reduce the independent variables as well as to help mitigate the issue of multicollinearity. The dimensions have been reduced from 6 to 3 principal components. Figure 4 shows that the first three PCs explain 89.55% of the variance. Their individual contributions are as follows: PC1 explains 61.71%, PC2 contributes 16.40%, and PC3 adds as much as 11.43%. With a cutoff of 0.5 for the loadings of components, as shown in Tables 2 and 3.

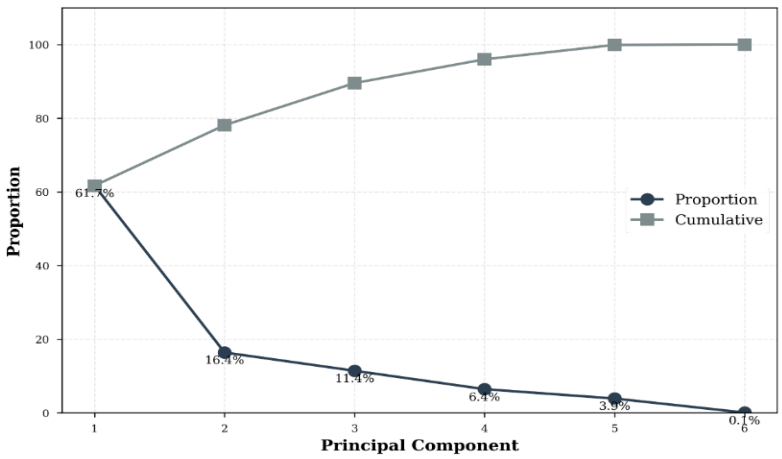

Figure 4. Proportion and Cumulative Variances Explained by the Produced PCS

The first principal component was associated mainly with osmolality and urea variables, with loadings of 0.5082 and 4709, respectively. The second component represented the pH variable, loading: 0.9123, and the third one reflected the calcium variable, loading: 0.8579. The generated component scores were then analyzed by logistic regression.

Table 2. Proportion and Cumulative Variances Explained by the Produced PCS.

Component	Eigenvalue	Variance Explained (%)	Cumulative Variance (%)
PC1	3.755	61.71	61.71
PC2	0.997	16.40	78.11
PC3	0.969	11.43	89.55
PC4	0.696	6.44	95.99
PC5	0.237	3.91	99.91
PC6	0.005	0.09	100.00

Table 3. Eigenvectors of the Considered PCS

Variable	PC1	PC2	PC3
Gravity	0.4596	-0.1373	-0.0959
pH	-0.1860	0.9123	0.1180
Osmo	0.5082	0.1312	-0.1526
Cond	0.3954	0.3554	-0.4616
Urea	0.4709	0.0008	0.0670
Calc	0.3433	0.0737	0.8579

Building on these results, a logistic regression model was performed using the first three principal components (PCS); the model fit statistics indicate that the model with covariates provides a better fit to the data compared to the intercept-only model. The lower values of AIC, SC, and -2 Log L for the model with covariates suggest improved model fit and a better trade-off between fit and complexity, as shown in Table 4.

Table 4. Model Fit Statistics of the Performed Logistic Regression

Criterion	Intercept	Intercept and Covariates
AIC	101.813	89.816
BIC	104.090	98.922
-2 Log L	99.813	81.816

These findings strongly suggest that the predictors play a meaningful role in the model and exhibit significant associations with the outcome variable. The global goodness-of-fit tests indicate that the fitted model is statistically significant. The Likelihood Ratio test ($\chi^2 = 17.99$, $p = 0.0004$), Score test ($\chi^2 = 24.44$, $p < 0.0001$), and Wald test ($\chi^2 = 20.87$, $p = 0.0001$) all reject the null hypothesis that the predictor coefficients are jointly equal to zero. These results shown in Table 5 demonstrate that the included explanatory variables collectively provide significant predictive power for the outcome and that the model offers a substantially better fit than an intercept-only model.

Table 5. Assessment of Beta Coefficients in Global Null Hypothesis Tests for Logistic Regression

Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio	17.99	3	0.0004
Score	24.44	3	<0.0001
Wald	20.87	3	0.0001

The maximum likelihood estimation presented in Table 6 indicates that the interception is not significant; hence, it plays little role in prediction as reflected by its estimate (estimate = 0.0735, p = 0.789). Prin1 is significant (estimate = 0.3339, p = 0.017), and we can say this variable has a critical effect on KSD risk.

Prin1 is statistically significant (estimate = 0.3339, p = 0.017), which means this variable has an important influence on kidney stone risk. Prin2 is not statistically significant (estimate = −0.1146, p = 0.693), suggesting that it does not contribute strongly to the prediction. In contrast, Prin3 shows a highly significant relationship with the outcome (estimate = 1.1147, p = 0.0001), making it the strongest predictor in the model. These results suggest that Prin1 and Prin3 have important roles, while the intercepts and Prin2 have little impact, which can be eliminated. In light of these results, the simplified logistic regression model can be expressed as

$$logit(p) = 0.3339 \cdot \text{Prin1} + 1.1147 \cdot \text{Prin3}$$

where logit(p) represents the log-odds of kidney stone risk.

Table 6. Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.0735	0.27	0.071	0.789
Prin1	1	0.3339	0.14	5.69	0.017
Prin2	1	-0.1146	0.29	0.15	0.693
Prin3	1	1.1147	0.29	14.47	0.0001

Table 7 presents correlations between observed outcomes and prediction probabilities, revealing potential support for predictive performance. The model had 75.5% concordant pairs, meaning that in roughly three-fourths of cases, it would rank patients by their relative level of risk correctly. The positive relationship is also indicated by the Somers' D and Gamma. The concordance statistic (c = 0.755) indicates that approximately 75.5% of the time the model may order risk predictions correctly, which is good discriminative performance. While 24.5% of pairings were discordant, the lack of tied pairs suggests that there are stable probability estimates between observations.

Overall, these logistic regression analyses suggest that osmolality (Prin1) and calcium concentration (Prin3) are important contributors to stone risk. Higher values of these variables are associated with increased predicted risk. Larger values of these variables are indicative of a greater predicted risk. Urine pH (Prin2), on the other hand, does not represent a statistically significant contribution to the model. These results emphasized the significance of biochemical concentration markers in predicting the risk of stone formation.

Table 7. Association of Predicted Probabilities and Observed Responses

Percent Concordant	75.5	Somers' D	0.511
Percent Discordant	24.5	Gamma	0.511
Percent Tied	0.0	Tau-a	0.259
Pairs	1296	c	0.755

4.3 Cluster Analysis

Hierarchical clustering has been used to determine the number of clusters. Both the elbow method graph and the dendrogram have been constructed as shown in Figures 5 and 6. This analysis determined that three clusters are ample for the dataset. while Figure 7 presents the clusters generated.

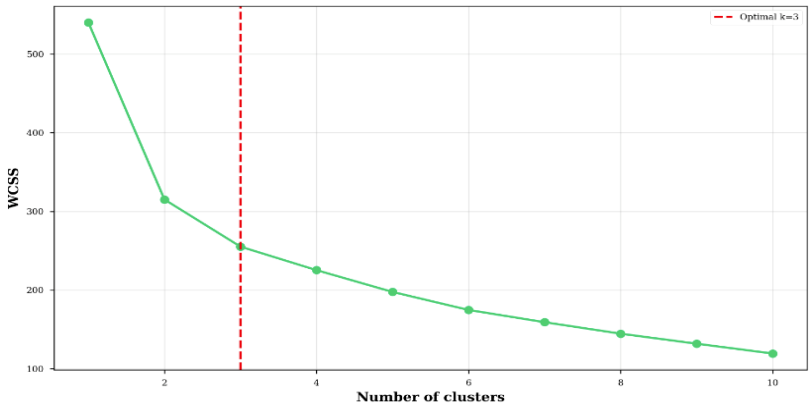

Figure 5. Elbow Method: Optimal Number of Clusters

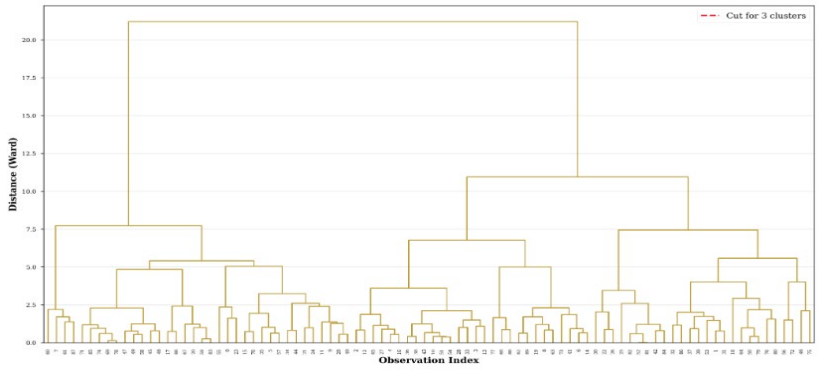

Figure 6. Hierarchical Clustering Dendrogram

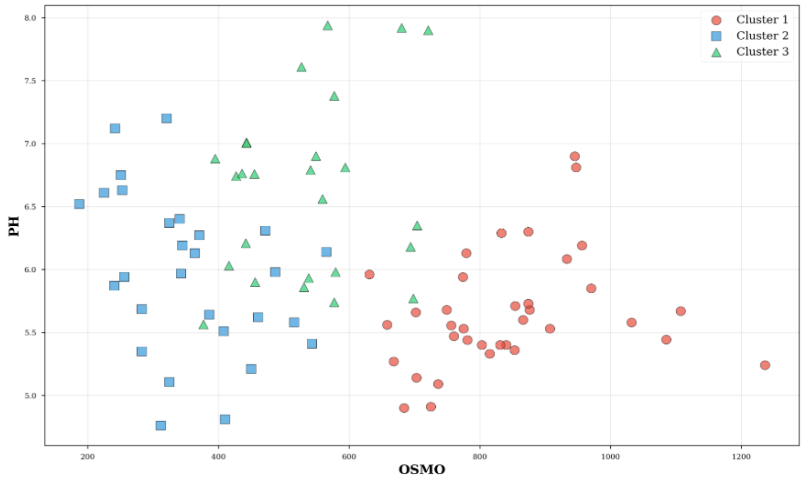

Figure 7. Clusters Scatter Plot

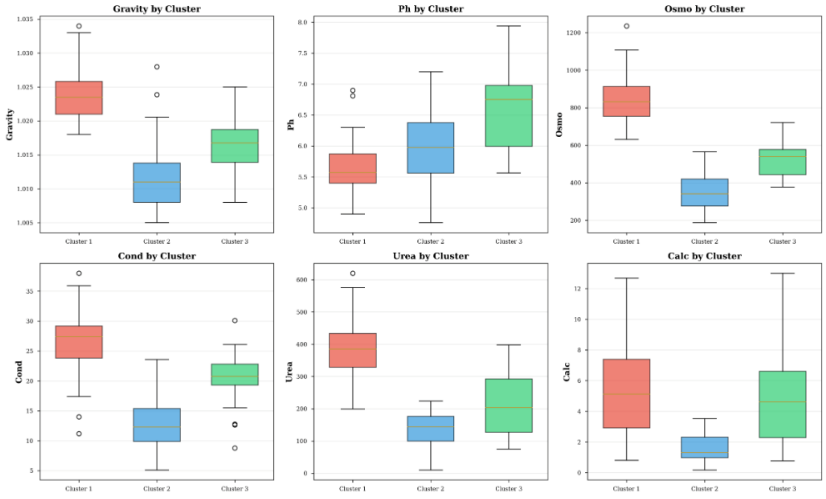

Figure 8. Clusters Profiles

In Figure 8, the first cluster observed had exceptionally higher values for osmolality, conductivity, urea, and calcium, which is an indication of concentrated urine with high mineral presence. This cluster showed a high kidney stone prevalence, which indicates that dehydration and higher solute concentration are the significant elements in lithiasis formation. In contrast, the second cluster demonstrated consistently low values across all laboratory measurements, including osmolality, conductivity, urea, and calcium. This cluster had the lowest stone prevalence, indicating that adequate hydration and reduced mineral levels contribute to a lower risk. The third cluster showed modestly higher serum calcium values and intermediate measurements in 24-hour urine concentration. Although the second cluster had a low osmolality compared to the first, there was an overall high prevalence of stone disease in this group, which emphasizes the important independent effect found for calcium concentration in kidney stones.

We extend the analysis in this article to search for the best method to achieve the optimal linkage, considering maximum performance. The cluster ability of clustering methods is evaluated using the results obtained from cluster analysis. The corresponding results are presented in Table 8.

Table 8. Hierarchical and KNN Clustering Results

Clustering Method	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
Ward Linkage	0.267	48.54	1.393
Average Linkage	0.328	40.70	1.044
Single Linkage	0.084	1.65	0.731
K-Means	0.267	51.73	1.369

5. Conclusion

The multivariate statistical methods, such as PCA, logistic regression, and hierarchical cluster analysis, were employed in this study to serve as strong predictive models for KSD formation. PCA reduced the six urinary biomarkers to 3 principal components, which accounted for 89.55% of the total variance in the data sets. The logistic regression model had good discriminatory power (c-statistic 0.755) to distinguish cases from non-cases (the probability that relative risks were correctly ordered was 75.5%). Overall goodness-of-fit tests showed that the model was highly significant (with a likelihood ratio = 17.99, $p = 0.0004$; and a Wald test statistic = 20.87, with $p = 0.0001$), strongly rejecting the null hypothesis that all coefficients are zero. Strongest predictors were identified as calcium-specific variation (Prin3, $\beta = 1.115$, $p = 0.0001$) and concentration in urine overall (Prin1, $\beta = 0.334$, $p = 0.017$) by examining coefficients of the significant strategies between two principal component factors (Fig J). Second, cluster analysis identified three distinct sets of patient phenotypes: patients in cluster 1 (55.6% stone prevalence) had highly

concentrated urine with high mineral content; those in cluster 2 (35.7% stone prevalence) exhibited low concentration and mineral content in the urine; and patients in cluster 3 (57.7% stone prevalence) showed the highest urinary calcium compared with relatively moderate overall concentration. Collectively, these data would suggest that multivariate statistical analysis can enhance differentiation between kidney stone risk factors and indicate specific biochemical targets for therapeutic intervention.

A limitation of this study is that the dataset was sourced from a public repository without documented clinical provenance; future work should validate these findings on datasets with verified collection protocols and larger sample sizes.

References

Abraham, A., Kavoussi, N. L., Sui, W., Bejan, C., Capra, J. A. and Hsi, R., Machine learning prediction of kidney stone composition using electronic health record-derived features, *Journal of Endourology*, vol. 36, no. 2, pp. 243–250, 2022.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. and Perona, I., An extensive comparative study of cluster validity indices, *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

Borghi, L., Schianchi, T., Meschi, T., Guerra, A., Allegri, F., Maggiore, U. and Novarini, A., Comparison of two diets for the prevention of recurrent stones in idiopathic hypercalciuria, *The New England Journal of Medicine*, vol. 346, no. 2, pp. 77–84, 2002.

Buch, A. M., Liston, C. and Grosenick, L., Simple and scalable algorithms for cluster-aware precision medicine, *Proceedings of Machine Learning Research*, vol. 238, pp. 136–144, 2024.

Doyle, P., Gong, W., Hsi, R. and Kavoussi, N., Machine learning models to predict kidney stone recurrence using 24 hour urine testing and electronic health record-derived features, *Research Square* (preprint), 2023, https://doi.org/10.21203/RS.3.RS-3107998/V1.

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X., *Applied Logistic Regression*, 3rd Edition, Wiley, 2013.

Husson, F., Josse, J. and Pagès, J., Principal component methods; hierarchical clustering; partitional clustering; why would we need to choose for visualizing data, Available: https://www.scirp.org/reference/referencespapers?referenceid=2950186, 2010.

Hyams, E. S. and Matlaga, B. R., Economic impact of urinary stones, *Translational Andrology and Urology*, vol. 3, no. 3, pp. 278–283, 2014.

Iparraguirre-Villanueva, O., Paucar-Palomino, G. and Paulino-Moreno, C., From data to diagnosis: Evaluation of machine learning models in predicting kidney stones, *Neural Computing and Applications*, vol. 37, no. 15, pp. 9049–9062, 2025.

Khalili, P., Jamali, Z., Sadeghi, T., Esmaeili-nadimi, A., Mohamadi, M., Moghadam-Ahmadi, A., Ayoobi, F. and Nazari, A., Risk factors of kidney stone disease: A cross-sectional study in the southeast of Iran, *BMC Urology*, vol. 21, no. 1, Article 141, 2021.

Mahmoodi, F., Andishgar, A., Mahmoudi, E., Monsef, A., Bazmi, S. and Tabrizi, R., Predicting symptomatic kidney stones using machine learning algorithms: Insights from the Fasa adults cohort study (FACS), *BMC Research Notes*, vol. 17, no. 1, Article 318, 2024.

Maugeri, A., Barchitta, M., Basile, G. and Agodi, A., Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions, *Scientific Reports*, vol. 11, no. 1, Article 4306, 2021.

Shastri, S., Patel, J., Sambandam, K. K. and Lederer, E. D., Kidney stone pathophysiology, evaluation and management: Core curriculum 2023, *American Journal of Kidney Diseases*, vol. 82, no. 5, pp. 617–634, 2023.

Stamatelou, K. and Goldfarb, D. S., Epidemiology of kidney stones, *Healthcare (Basel)*, vol. 11, no. 3, Article 424, 2023.

Teichman, J. M. H., Clinical practice: Acute renal colic from ureteral calculus, *The New England Journal of Medicine*, vol. 350, no. 7, pp. 684–693, 2004.

Ye, W., Lu, W., Tang, Y., Chen, G., Li, X., Ji, C., Hou, M., Zeng, G., Lan, X., Wang, Y., Deng, X., Cai, Y., Huang, H. and Yang, L., Identification of COVID-19 clinical phenotypes by principal component analysis-based cluster analysis, *Frontiers in Medicine*, vol. 7, Article 570614, 2020.

Biographies

Omar Abueed is a PhD candidate in Systems Science and Industrial Engineering at Binghamton University, specializing in artificial intelligence, deep learning, and healthcare analytics, with research focused on medical image segmentation and AI-enabled decision support.

Audai Al-Akailah is a Ph.D. researcher in mechanical engineering at Binghamton University (SUNY). His research focuses on advanced manufacturing processes, with particular emphasis on modeling, process characterization, and predictive understanding of complex manufacturing systems. His work combines experimental investigation with computational and statistical methods to study process–structure–property relationships. In addition to manufacturing applications, his research interests extend to interdisciplinary studies involving predictive modeling in biomedical contexts.

Maya Abuarayes is a Master student in Industrial and Systems Engineering at the State University of New York, Binghamton University. Maya's research interests include healthcare data mining, prediction models, machine learning, artificial intelligence, large language models (LLMs), and lean continuous improvement methodologies.

Mohammed Abueed holds a Ph.D. in Industrial and Systems Engineering from Auburn University, where his research focused on the reliability modeling of lead-free solder joints under fatigue, creep, and thermal cycling conditions. He currently serves as a Module Development Engineer at Intel Corporation. His work spans electronics reliability, predictive modeling, failure analysis, and continuous improvement, and he has authored numerous publications in IEEE, SMTA, and other leading venues.

Shane Menezes is a PhD candidate in Systems Science at Binghamton University (SUNY). His research focuses on the reliability, resilience, and optimization of the New York State energy grid. His work combines modeling, simulation, sensitivity analysis, operations research, and network science. His research interests extend to net-zero carbon policies and future uncertainties, including the impacts of extreme weather events.

Yong Wang is an associate professor and associate school director in the School of Systems Science and Industrial Engineering at Binghamton University, New York, USA. His research interests include autonomous systems, energy systems, manufacturing systems, healthcare systems, operations research, and data science.