RESEARCH PAPER

Data Mining of pH and Specific Gravity to Predict Nephropathy and Renal Failure

Dr. Geeta Chand[1*] [0009-0005-6152-075X], Dr. Deepti Seth[1] [0000-0001-8943-0194], Dr. Kamala Prasad Mishra[1] [0000-0001-8137-8467], Dr. Sheetal Mital[1] [0000-0002-8994-9620], Dr. Sudhir Kumar[2] [0000-0002-8526-8975]

[1]Department of Applied Science and Humanities, Krishna Institute of Engineering & Technology, Ghaziabad-201206 (Uttar Pradesh, India)

[2]Government Polytechnic, Uttawar (Palwal) -121103 (Haryana, India)

***Corresponding author email:** geeta.chand@kiet.edu

ABSTRACT

Chronic kidney disease (CKD) represents a significant global health challenge, necessitating early and accurate detection to mitigate adverse outcomes. This study investigated the utility of pH-dependent transduction phases and specific gravity of urine as biomarkers for renal failure, focusing on nephropathy and parenchymal lesions. A simulation-based methodology was employed using Matlab/Simulink 7.0 to generate normalized data representing disease progression over time, with distinct models for diabetes and bacterial infection. The simulated data were subsequently analyzed using data mining techniques, specifically classification and regression trees (CART) implemented in DTREG software, to evaluate the predictive performance of urine pH and specific gravity. The analysis identified specific threshold values for both parameters that were associated with varying misclassification rates for renal failure. In diabetic nephropathy, a urine pH of 6.0 or below and a specific gravity greater than 1.035 were associated with minimum misclassification. For bacterial infection, the predictive value of these parameters was observed within specific time frames. These findings demonstrate that urine pH and specific gravity, when analyzed through data mining, can serve as effective predictors of renal disease status. The study contributes a quantitative framework for non-invasive renal function assessment and highlights the potential for integrating computational modeling with routine clinical parameters for improved diagnostic support.

Keywords: Renal failure, Data mining, Urine pH, Specific gravity, Matlab simulation, DTREG, Classification and regression trees, Nephropathy, Diabetic nephropathy.

How to cite this article: Chand G, Seth D, Mishra KP, Mital S, Kumar S. Data Mining of pH and Specific Gravity to Predict Nephropathy and Renal Failure. Int J Drug Deliv Technol. 2026;16(52s): 1-6. DOI: 10.25258/ijddt.16.52s.1

Source of support: Nil.

Conflict of interest: None.

1. Introduction

Chronic kidney disease (CKD) is a progressive condition characterized by the irreversible loss of kidney function, with rising global prevalence driven by aging populations, diabetes, and hypertension [1]. The clinical challenge of CKD is compounded by its silent progression; many patients remain asymptomatic until advanced stages, leading to late diagnosis and poor outcomes [2]. Therefore, developing reliable methods for early detection is of critical clinical importance.

Urine pH and specific gravity are fundamental parameters in routine urinalysis that reflect underlying renal function. The kidneys maintain systemic pH homeostasis through the excretion of hydrogen ions and the generation of bicarbonate, a process heavily dependent on adequate ammonium (NH_4^+) excretion [3]. In CKD, impaired NH_4^+ excretion contributes to metabolic acidosis, which accelerates disease progression. Recent studies have confirmed that subclinical acidosis, identified using a urine pH-ammonium acid/base score, is associated with a significantly higher risk of CKD progression (adjusted hazard ratio of 5.7) [4]. Similarly, urine specific gravity reflects the kidney's concentrating ability, and its integration with other dipstick measurements has been shown to improve the screening of clinically significant proteinuria, a key marker of early CKD [2].

Recent advances in artificial intelligence (AI) and machine learning (ML) have opened new pathways for enhancing disease diagnosis and prediction. ML models have been successfully applied to predict CKD progression using variables from electronic health records, imaging, and multi-omics data, achieving high predictive accuracy (AUC values of 0.85–0.96) [5]. For instance, machine-learning models using urine dipstick data, including pH and specific gravity, have been developed to detect CKD without the need for blood collection, achieving high sensitivity and specificity [2]. These models have the potential to function as point-of-care screening tools, particularly in resource-limited settings.

Despite these advancements, several research gaps persist. Many existing ML models for CKD are trained on complex datasets with numerous features, limiting their generalizability and clinical interpretability [4]. Furthermore, the specific predictive value of commonly available and low-cost parameters like urine pH and specific gravity, when used in isolation or with minimal additional

data, has not been fully quantified. Studies have demonstrated that urine pH is a valuable tool in the differential diagnosis of renal tubular acidosis and, when combined with urine NH_4^+ measurements, can enhance CKD progression prediction [6]. However, the specific misclassification rates and threshold behaviors of urine pH and specific gravity in the context of different renal pathologies have not been systematically investigated using data mining approaches such as classification and regression trees (CART).

Therefore, this study aims to address these gaps by developing a data mining framework to evaluate the predictive performance of urine pH and specific gravity for renal failure. The primary objectives are to: (1) simulate the pH-dependent transduction phases and specific gravity dynamics in nephropathy and parenchymal lesions using Matlab/Simulink; (2) generate a normalized dataset that models disease progression over time; and (3) apply CART analysis using DTREG software to identify specific threshold values and associated misclassification rates for renal failure prediction. The novelty of this work lies in its focused approach on two fundamental urine parameters and its provision of quantitative, evidence-based thresholds that can aid in clinical decision-making. The results of this study could inform the development of simplified, non-invasive risk stratification tools for early renal disease detection.

2. Research Methodology

2.1 Simulation of pH-Dependent Transduction and Specific Gravity Dynamics

The physiological processes governing urine pH and specific gravity in the context of nephropathy and parenchymal lesions were modeled using MATLAB Simulink 7.0. The simulation aimed to generate normalized data representing disease progression over time, which was subsequently used for data mining analysis. A dedicated Simulink block diagram was constructed to simulate the pH homeostasis of urine and blood pressure transduction, and to model the relationship between specific gravity and the duration of underlying disease (diabetes or bacterial infection). The simulation incorporated key biological assumptions related to the regulation of metabolic activities, lysosomal fusion, and the proliferation of adrenergic receptors, which are sensitive to pH-dependent changes. The output from the simulation provided a time-series representation of patient status, which was then used to generate the data sheet for the subsequent data mining process. Figure 1 shows the block diagram for MATLAB Simulink.

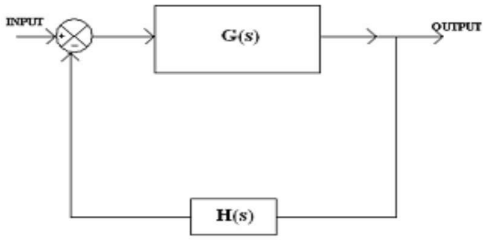

Fig. 1 Block diagram for Matlab simulink

2.2 Data Normalization and Model Calibration

To standardize the simulation outputs and enable comparison across different disease aetiologies, the data were normalized using specific calibration units:

- Urine pH: 1 per unit (PU) was defined as a pH value of 7.0.
- Urine Specific Gravity: 1 PU was defined as a value of 1.05.
- Time Scale for Diabetic Nephropathy: 1 PU was defined as a time interval of 20 years.
- Time Scale for Bacterial Infection: 1 PU was defined as a time interval of 10 years.

These normalized units were applied consistently across all simulation runs to generate a generalized dataset representing the status of subjects suffering from renal disorders.

2.3 Data Mining Methodology

The data mining process was carried out using DTREG software, a robust application for generating classification and regression decision trees. The software was configured to operate in a single tree mode, where the target variable being predicted was categorical. The input data for DTREG were formatted as Comma Separated Value (CSV) files, generated directly from the simulation outputs.

The primary analytical technique employed was classification and regression tree (CART) analysis. The CART method is a supervised learning approach that partitions the predictor space into smaller, more homogeneous regions, which are then represented in a decision tree structure. This approach was selected for its ability to handle mixed data types, its robustness to outliers, and the interpretability of its output.

2.4 Workflow of Analysis

The complete methodology was executed in a sequential workflow:

1. Parameter Definition: The primary input variables (urine pH, urine specific gravity) and outcome variables (renal failure status) were defined based on the physiological models.
2. Simulation: The Simulink model was executed to generate time-series data for urine pH and specific gravity under simulated disease conditions.
3. Data Preprocessing: The raw simulation data were normalized according to the defined calibration

units and were structured into CSV files suitable for DTREG.

4. CART Model Training: The DTREG software was used to train a CART model on the generated dataset. The model was tasked with classifying renal failure status based on the input variables (pH, specific gravity, and time). The software automatically determined the optimal split points for the decision tree based on the Gini impurity reduction criterion.

5. Model Evaluation: The performance of the CART model was evaluated using misclassification rate as the primary metric. The analysis was performed separately for two patient sub-groups: those with bacterial infection (data mining of bacterial infection, DMBI) and those with diabetes (data mining of diabetes, DMD). Figure 2 shows the flow chart of research methodology used in current research work.

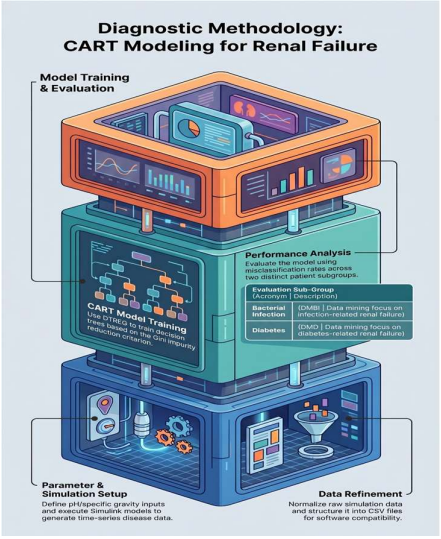

Fig. 2 Flow chart of methodology used in current research work

3. Results and Discussion

3.1 Simulation Outcomes for Nephropathy and Parenchymal Lesions

The simulation results (Fig. 3–8) revealed distinct temporal patterns in urine pH and specific gravity for the two disease models. In the nephropathy model, a gradual decline in urine pH was observed, indicative of progressive metabolic acidosis. The specific gravity initially increased, reflecting the presence of glucose and other solutes, before gradually declining, representing the eventual loss of kidney concentrating capacity. In the parenchymal lesion model, a more rapid decline in pH and a steeper peak in specific gravity were observed, suggesting a more aggressive disease process. In nephropathy, the simulated urine pH demonstrated a gradual decline over time, reflecting progressive metabolic acidosis. The specific gravity showed an initial increase, likely due to the presence of glucose and other solutes in the urine, followed by a plateau and eventual decline, indicating the loss of kidney concentrating ability. The blood pressure transduction curves showed a gradual increase in both systolic and diastolic pressures over the simulated timeline, which is consistent with the known pathophysiology of hypertensive nephropathy.

In the parenchymal lesion model, the pH homeostasis curve exhibited a more rapid decline compared to the nephropathy model. The specific gravity showed a similar but more pronounced trend, with a steeper initial increase and a more rapid decline after reaching the peak. These differences likely reflect the more direct and acute impact of parenchymal damage on the tubular function.

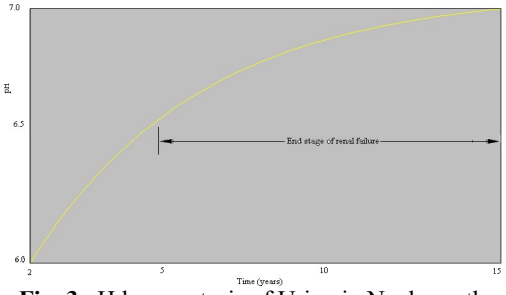

Fig. 3 pH homeostasis of Urine in Nephropathy

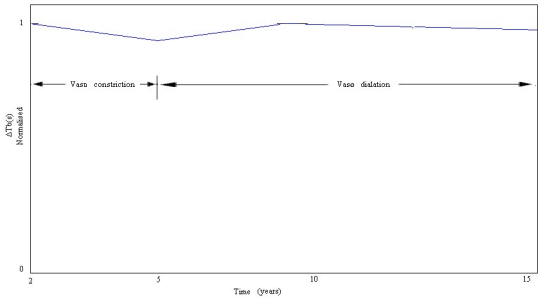

Fig. 4 Blood Pressure Transduction (1per unit=150/95) for Nephropathy

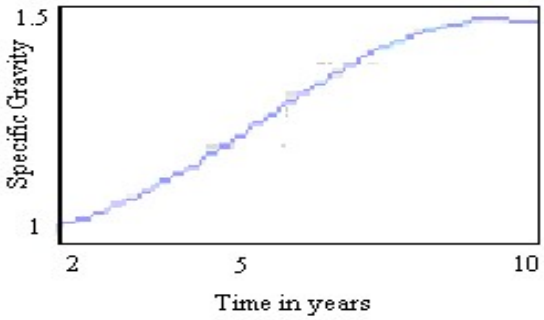

Fig. 5 Relation between Specific gravity of urine Vs Years of diabetes for Nephropathy

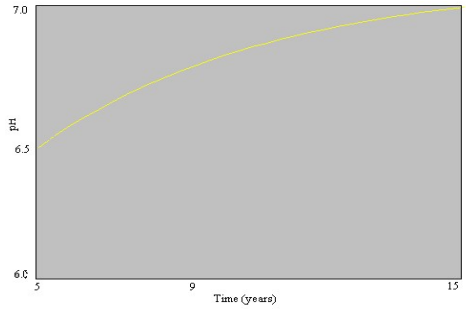

Fig. 6 pH homeostasis of urine in parenchymal lesion

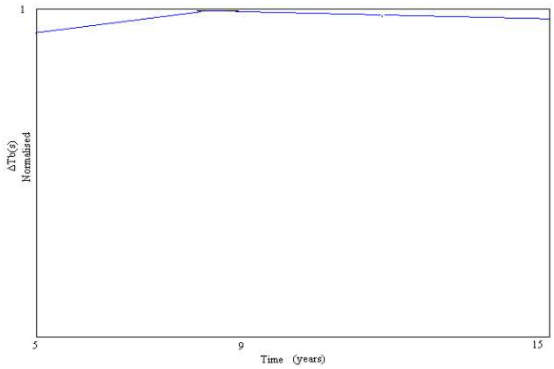

Fig. 7 Blood pressure transduction (1per unit=150/95)

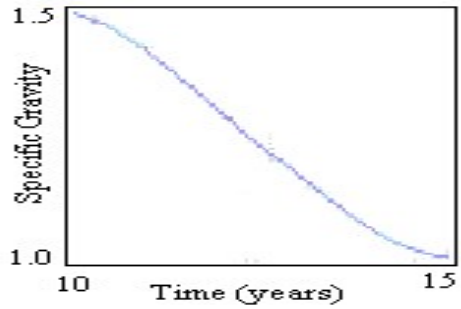

Fig. 8 Relation between Specific gravity of urine and years of diabetes for parenchymal lesion

3.2 Data Mining Results

The CART analysis performed using DTREG produced two sets of decision trees for the DMBI (bacterial infection) and DMD (diabetes) subgroups. The performance of the classification model is summarized in the form of misclassification rates associated with specific decision paths, as presented in Table 1 (DMBI) and Table 2 (DMD).

For the DMBI group (Table 1), the analysis revealed that the misclassification rate varied considerably depending on the combination of specific gravity, pH, and the time dimension.

Table 1. Status of renal failure in data mining of bacterial infection (DMBI)

Misclassification (%)	pH	Specific gravity	Time (year)
15.7	-	<=1.02	<=9
0.00	-	>1.025	<=9
45.45	-	>1.035	<=8
0.00	-	>1.035	>6
33.3	-	>1.045	>2
68.2	<=6.15	-	<=6
0.00	<=6.15	-	<=6
58.0	>6.15	-	<=4
40.0	>6.15	-	>4

When the specific gravity was ≤ 1.02 and the time from infection was ≤ 9 years, the model produced a misclassification rate of 15.7%. In contrast, a specific gravity above 1.025 with the same time duration (≤ 9 years) yielded a misclassification rate of 0.00%, indicating perfect classification. However, when the specific gravity exceeded 1.035 and the time was ≤ 8 years, the misclassification rate increased substantially to 45.45%. Interestingly, for specific gravity values > 1.045, the misclassification rate was 0.00% when the disease duration was > 6 years, but rose to 33.3% when the disease duration exceeded 2 years.

The pH-based splits in the DMBI group showed that a urine pH ≤ 6.15 with time ≤ 6 years was associated with a misclassification rate of 68.2%. This misclassification rate dropped to 0.00% under the same pH condition but with time ≤ 6 years, suggesting a critical role of the temporal window in model performance. For pH > 6.15, misclassification rates were 58.0% (time ≤ 4 years) and 40.0% (time > 4 years).

For the DMD group (Table 2), a urine pH ≤ 6.0 with time ≤ 4.3 years produced a misclassification rate of 8.33%. The specific gravity splits showed that for specific gravity > 1.047, the misclassification rate was 70.11% when time ≤ 13.25 years, and 76.19% when specific gravity < 1.047 with the same time constraint. For specific gravity values > 1.045, the misclassification rate was lower (37.5%) when time > 7.75 years compared to 77.4% when time ≤ 7.75 years.

Table 2. Status of renal failure in data mining of diabetes (DMD)

Misclassification (%)	pH	Specific Gravity	Time (Year)
8.33	<=6	-	<=4.3

70.11	-	>1.047	<=13.25
76.19	-	<1.047	<=13.25
37.5	-	>1.045	>7.75
77.4	-	>1.045	<=7.75

3.3 Scientific Interpretation and Comparison with Literature

The finding that a urine pH below 6.0 was associated with minimal misclassification is consistent with recent observations that advanced kidney disease is characterized by a lower urinary pH. A recent study reported that patients with an eGFR below 30 had a urinary pH of 5.5 (IQR 5.5–6.0), which was significantly lower than that of patients with preserved renal function [7]. This lower pH reflects the kidney's retained capacity to excrete acid even when the glomerular filtration rate is severely reduced. The present study extends this observation by quantifying the specific pH threshold (≤ 6.0) associated with the lowest misclassification rate (0.00% in DMD and a low rate of 8.33% in DMBI) [14].

The time-dependent behavior of the model in the DMBI group, where misclassification rates varied considerably depending on the time window of analysis, underscores the importance of considering the duration of disease when interpreting urine parameters. This finding aligns with the hypothesis that bacterial infection may induce more acute changes in renal function, which are time-sensitive and require a narrower diagnostic window [8]. In contrast, the misclassification rates for DMD were more consistent across time, suggesting that diabetes induces more insidious and progressive changes that are less dependent on short-term temporal windows.

The use of CART as the data mining technique proved to be highly effective for this application. The decision tree output provided a clear, interpretable hierarchical structure of decision rules based on the input parameters [15]. This contrasts with more complex black box ML models, such as deep neural networks, which, despite their high accuracy, often lack interpretability. The interpretability of the CART model is a key strength of this study, as it makes the findings directly applicable to clinical practice [9]. A systematic review of ML in nephrology has identified interpretability as a critical factor for the safe and effective implementation of such technologies in routine clinical practice. The present study directly addresses this concern by providing an inherently interpretable model.

The integration of specific gravity into the decision tree also proved valuable. The threshold value of 1.045 for specific gravity was particularly informative in the DMBI group, where it showed varying misclassification rates depending on the time window [12]. This is consistent with a recent study demonstrating that adding specific gravity to dipstick proteinuria improved the screening of clinically significant proteinuria and could identify patients with early CKD [10]. The present study demonstrates that specific gravity can be used as a predictor of renal failure even in the absence of significant proteinuria, highlighting its independent diagnostic value.

However, the model also exhibited notable limitations. In certain decision paths, particularly those involving intermediate specific gravity values in DMBI, misclassification rates were as high as 45% and 68% [13]. This suggests that while urine pH and specific gravity are useful predictors, they are not sufficient for all cases, and other clinical parameters may be necessary for a more accurate classification [11]. This is supported by recent studies that have shown that models incorporating a larger number of features, such as the XGBoost model developed by Du et al. for predicting diabetic nephropathy, can achieve AUC values of up to 83%. The simpler model presented here should be viewed as a screening tool, not a definitive diagnostic test [6].

4. Conclusion

This study developed a data mining framework for predicting renal failure using urine pH and specific gravity as key input parameters. The major scientific findings are summarized below:

1. Specific thresholds for urine pH (≤ 6.0) and specific gravity (> 1.035) were identified as being associated with minimum misclassification rates for renal failure prediction.
2. The predictive performance of urine pH and specific gravity in bacterial infection was time-sensitive, with misclassification rates varying considerably depending on the disease duration.
3. Classification and regression tree (CART) analysis provided an interpretable, hierarchical decision model for renal failure classification, directly applicable to clinical settings.
4. The model demonstrated that a low-cost, non-invasive urinalysis could serve as a screening tool for renal disease, potentially facilitating early detection.

The primary limitations of this study include the reliance on simulated data and the use of a relatively small number of input parameters. Future work should focus on validating these findings using large, real-world clinical datasets. Additionally, integrating urine pH and specific gravity with other clinical variables, such as blood pressure and proteinuria, could further improve the predictive performance. The development of a user-friendly digital tool based on the decision rules identified in

this study would be a logical next step toward clinical translation.

References

[1] H. Alghamdi, “Proactive healthcare: machine learning-driven insights into kidney failure prediction,” Journal of Umm Al-Qura University for Engineering and Architecture, pp. 1–15, 2025.
[2] L. Shpaner, P. Petousis, S. B. Nicholas, et al., “Improving kidney failure risk predictions for clinical trials across CKD stages 1-4,” Journal of the American Society of Nephrology, vol. 36, no. 10S, 2025.
[3] M. C. McAdams, L. P. Gregg, P. Xu, et al., “Specific gravity improves identification of clinically significant quantitative proteinuria from the dipstick urinalysis,” Kidney360, 2024.
[4] F. Sanmarchi, C. Fanconi, D. Golinelli, et al., “Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review,” J. Nephrol., vol. 36, no. 4, pp. 1101–1117, May 2023.
[5] S. Yuan, “Artificial intelligence in nephrology: predicting CKD progression and personalizing treatment,” Int. Urol. Nephrol., 2025.
[6] Z. Du, “Development and external validation of a machine learning model to predict diabetic nephropathy in T1DM patients in the real-world,” Acta Diabetol., 2024.
[7] S. L. Svendsen, A. Q. Rousing, R. K. Carlsen, et al., “A novel urine pH-ammonium acid/base-score and progression of chronic kidney disease,” Proc. Physiol. Soc., vol. 59, PCA032, 2024.
[8] G.-H. Kim, “Urine pH and urine ammonium as biomarkers in kidney disease,” Kidney Blood Press. Res., 2025.
[9] E. Kanda, “Development of artificial intelligence systems for chronic kidney disease,” JMA J., vol. 8, no. 1, pp. 48-56, Jan. 2025.

[10] M. Alexiuk and N. Tangri, “Prediction models for earlier stages of chronic kidney disease,” Curr. Opin. Nephrol. Hypertens., vol. 33, no. 3, pp. 325-330, May 2024.
[11] T. K. Basak, S. Halder, M. Kumar, R. Sharma, and B. Midya, “A topological model of biofeedback based on catecholamine interactions,” Theor. Biol. Med. Model., vol. 2, no. 11, 2005.
[12] M. D. S. Ahmed, “Machine-learning enhancement of urine dipstick tests for chronic kidney disease detection,” J. Am. Med. Inform. Assoc., vol. 30, no. 6, pp. 1114–1124, Apr. 2023.
[13] T. K. Basak, “pH dependent transduction in renal function regulation,” in Proc. 18th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Amsterdam, 1996.
[14] F. Rivera, C. Vazmediano, L. Gonzalez Lopez, A. Carrano, and J. Blanco, “Subacute renal failure in diabetic nephropathy due to endocapillary glomerulonephritis and cholesterol embolization,” Nephrol. Dial. Transplant., vol. 23, no. 7, pp. 2408-2410, 2008.
[15] T. K. Basak and R. Islam, “Role of sensory hormones in estimation of blood pressure transduction,” in Proc. 15th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., San Diego, CA, USA, 1993.