Received Date:1st December, 2025
Revision Date: 24th December, 2025
Accepted Date: 27th February, 2026

Early Prediction of Chronic Kidney Disease Using Extra Trees and Light GBM with Sharp Visualization

Sujan Gupta[1*] & Rajad Shakya[2]

[1]Dept of Electronics and Computer Engineering, Thapathali Campus, TU, Nepal, E-mail: sujangupta217@gmail.com
[2]Lecturer, Dept of Electronics and Computer Engineering, Thapathali Campus, TU, Nepal, E-mail: shakyarajad1@gmail.com

Abstract— This research paper aims to detect Chronic Kidney Disease (CKD) at the initial stages by using blood and urine test parameters leveraging computationally efficient machine learning algorithms such as LightGBM and Extra Trees. The UCI Chronic Kidney Disease dataset with 400 instances and 24 features along with binary "ckd" and "notckd" target class, is preprocessed by imputing missing values, clipping outliers and normalization. Validation was done by 5-fold cross validation technique. Extra Trees and LightGBM achieved accuracies of 0.993 and 0.985 respectively. SHAP visualization showed albumin, hemoglobin and specific gravity (sg = 1.025) as key features using Extra Trees whereas LightGBM showed albumin, hemoglobin and serum creatinine as key features. Albumin, hemoglobin, specific gravity = 1.025, hypertension and serum creatinine are significant indicators of CKD. Specific value of sg i.e. 1.025 being a significant contributor can be a significant area of medical study. Hence, this study offers an interpretable, lightweight framework suitable for integration into routine clinical blood testing for early CKD detection and can be emulated as a method of identifying features that contribute towards a disease, which with high model accuracy can be put into consideration for clinical research to uncover the feature's biological meaning and potential as a hidden biomarker as well. Notably, low hemoglobin, linked to reduced production of erythropoietin as seen in impaired kidneys, strongly predicted "ckd", displaying one such critical biomarker. This shows a strong potential of using this method for other diseases as well.

Keywords— Chronic Kidney Disease, Explainable AI, Extra Trees, Interpretability, LightGBM, Machine Learning, Medical Diagnosis, SHAP, UCI Repository

Introduction

Chronic Kidney Disease (CKD) is a long-term disorder where kidneys in a person's body gradually lose their ability to function properly, damaging their waste and liquid filtering ability. It is a global health issue that remains largely a case of asymptomatic disorder not getting detected early on, when the chance for treatment is highly feasible. The main area of concern is the production of methods to accurately detect CKD using minor tests or unrelated blood and urine tests for a patient, since regular periodic checkups just for possibility of hidden diseases is unfeasible.

Conventional medical approaches for CKD involve expensive and tedious tests which make routine screening unfeasible. Recent studies have been conducted using Random Forest, XGBoost, and Artificial Neural Networks (ANN) achieving up to 97.5% accuracy using ANN [1] using blood and urine test data. However, these approaches are often computationally expensive and lack interpretability.

The main objectives of this study are as follow:

- This study aims to apply Extra Trees [2], a variation of random forests [3] and LightGBM [4], a gradient boosting framework for early and efficient CKD detection.
- This study aims to apply SHAP [5] visualization technique to interpret the predictions made and study the impact of features towards the prediction.

In conclusion, in this study, we apply Extra Trees and LightGBM models to predict CKD from the UCI Repository dataset [6] and assess their performance and interpretability using SHAP.

Related Works

Poudel [1] investigated the performance of the same UCI dataset across three different models namely Random Forest, XGBoost and ANN. Though he achieved the highest accuracy of 97.5% using ANN, it was computationally not as efficient as statistical modern machine learning algorithms such as Extra Trees and LightGBM. Extra Trees proposed by Geurts et al enhanced Random Forests [2] with randomized features and split selection [3]. LightGBM [4], introduced

* Corresponding Author

by Ke et al., uses histogram-based learning offering better results than XGBoost. SHAP [5], by Lundberg and Lee, addresses interpretability by sorting feature importance.

Ghosh and Khandoker [7] proposed a similar paper using five machine learning (ML) methods, logistic regression (LR), random forest (RF), decision tree (DT), Naïve Bayes (NB) and XGBoost, similarly SHAP visualization in the paper predicted creatinine, glycosylated hemoglobin type A1C (HgbA1C), and age being the significant indicators. However, the highest achieved accuracy was only 93.29%. The uses of five different models with varying performance makes clinical integration difficult. Similarly, this study fails to discover unique insights and the predictions generated (age, sugar level and creatinine) are generic. This study uses a clinical dataset which is less accessible.

In contrast, this study aims to develop a computationally efficient and lightweight framework using only ExtraTree and LightGBM, both of which offer better accuracy and interpretability in SHAP plots, suitable for clinical integrations. The predicted features such as creatinine, hemoglobin, albumin and specific gravity can be considered as more accurate in developing medical insights due to increased classification ability of the models i.e. higher accuracy. The UCI dataset used in the study is publicly available with standard blood and test parameters suitable for reproducibility. Similarly, specific feature values such as (specific gravity = 1.025) offer novel insights for use in medical research. Furthermore, this study pushes this unique method of generating important features that should be put into consideration for medical research based on important features generated by any interpretability methods such as SHAP with a model having high enough accuracy to develop critical and hidden biomarkers.

Related Theory

Extra Trees (Extremely Randomized Trees) builds upon the Random Forest model by introducing randomization in both feature selection and split thresholds, decreasing overfitting, making it suitable for high-dimensional, noisy datasets such as the UCI CKD dataset. This algorithm is a robust ensemble method that constructs multiple decision trees by recursively dividing the dataset based on feature thresholds. LightGBM, a gradient boosting method, optimizes training through histogram-based learning and leaf-wise tree growth, increasing performance parameters as compared to XGBoost. SHAP (SHapley Additive exPlanations) uses cooperative game theory to quantify the contribution of each feature to model predictions, offering interpretability and visualisation. These methodologies collectively enable accurate, efficient, and interpretable CKD classification, suitable for clinical applications.

Dataset Overview

4.1. Dataset Description

The dataset for this study was obtained from the UCI Machine Learning Repository, specifically the Chronic Kidney disease dataset. This dataset contains 400 instances with 24 attributes/features including patient demographics, clinical measurements, and laboratory results related to CKD diagnosis, such as age, blood pressure, serum creatinine level, and urine albumin level.

Table 1.

List of all features in UCI Dataset with their datatypes

Features	Type	Description	Missing Values
age	Integer	Age	yes
bp	Integer	blood pressure	yes
sg	Categorical	specific gravity	yes
al	Categorical	albumin	yes
su	Categorical	sugar	yes
rbc	Binary	red blood cells	yes
pc	Binary	pus cell	yes
pcc	Binary	pus cell clumps	yes
ba	Binary	bacteria	yes
bgr	Integer	blood glucose random	yes
bu	Integer	blood urea	yes
sc	Continuous	serum creatinine	yes
sod	Integer	sodium	yes
pot	Continuous	potassium	yes
hemo	Continuous	hemoglobin	yes
pcv	Integer	packed cell volume	yes
wbcc	Integer	white blood cell count	yes
rbcc	Continuous	red blood cell count	yes
htn	Binary	hypertension	yes
dm	Binary	diabetes mellitus	yes
cad	Binary	coronary artery disease	yes
appet	Binary	appetite	yes
pe	Binary	pedal edema	yes
ane	Binary	anemia	yes
class	Binary	ckd or not ckd	no

4.2. Data Cleaning

The dataset for this study was imported from the UCI Machine Learning Repository, specifically the Chronic Kidney disease dataset. This dataset consists of tab '\t' strings in dm and class columns. These need to be removed firstly before any other operations or else one hot encoding will create a separate column for name_\t which the model will consider as a relevant feature even though it has no real world meaning.

- dm : row index=188=’\t’
- class: row index=37 & 230=’\t’

Similarly, the dataset consists of numerous outliers and noise which may disturb the model accuracy. These outliers are clipped using upper and lower bounds.

$$Upper\ Bound = Q3 + 1.5 * (Q3 - Q1) \quad \text{(Equation 1)}$$

$$Lower\ Bound = Q1 - 1.5 * (Q3 - Q1) \quad \text{(Equation 2)}$$

Where Q1 and Q3 are first and third quartiles

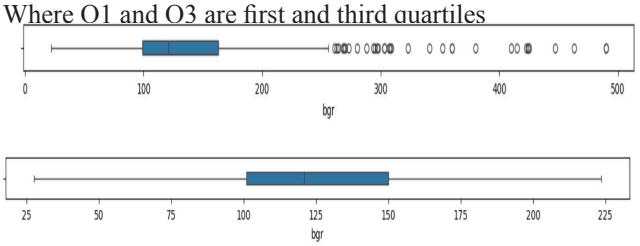

Figure 1. Outliers (white circles) in blood glucose random (bgr) before (a) and after (b) Clipping

4.3. Data Encoding

This dataset has categorical and binary features such as the albumin, specific gravity, ckd. Such values are incomprehensible for a computer and the model so they are encoded. The categorical classes are encoded using one hot encoding, similarly, the binary target “class” features are encoded using label encoding with ckd=1 and notckd=0

4.4. Data Normalisation

The integer based features have disproportionate ranges with some ranging from 0 to 1 to some from 10000 to 50000 hence we need to normalise the range to reduce computation cost as well as removing unwanted bias. For this standard scalar is used which is:

$$z = \frac{x - \mu}{\sigma} \quad \text{(Equation 3)}$$

4.5. Dataset Splitting

The preprocessed dataset is then split into an 80:20 ratio with 80% data for training pairs and 20% for testing/validation pairs. The split dataset is shuffled randomly to eliminate biases.

I. Methodology

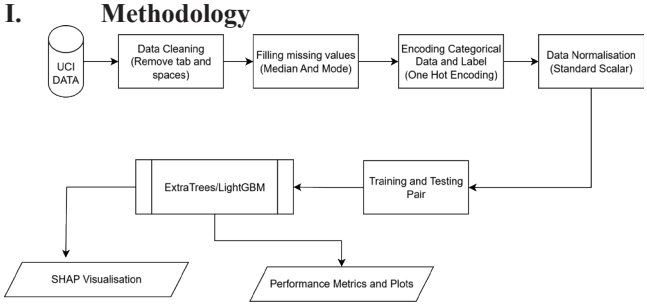

Figure 2. Working Mechanism of CKD prediction model

The methodology encompasses the following steps:

5.1. Data Preprocessing

The dataset was filled using mode and median for categorical and numerical features respectively, cleaned by removing erroneous tab characters, encoded using label and one hot encodings for categorical data, normalized using Standard Scaling then finally splitted in 80:20 ratio for training and testing.

5.2. Model Training

5.2.1. Model Configuration

Table 2.

Model Configuration for Extra Trees and LightGBM

Parameter	Extra Trees	LightGBM
Algorithm	ExtraTreesClassifier	LGBMClassifier
Number of Estimators	100	100
Random State	42	42
Max Depth	10	5
Learning Rate	Not Applicable	0.1
Training Data Split	80:20 Train:Test Data	80:20 Train:Test Data

5.3. Evaluation

Models were evaluated using 5-fold cross-validation, with parameters such as accuracy, precision, recall, and F1 score. Confusion matrices were generated for both the models.

5.4. Interpretability

5.4.1. SHAP Theory

SHAP (SHapley Additive exPlanations) uses Shapley values from cooperative game theory to quantify each feature’s contribution to CKD predictions, ensuring fair attribution across all feature combinations [5]. It provides additive feature attributions, expressed as:

$$g(z) = \varphi 0 + \sum_{i=1}^{M} \varphi i * zi \quad \text{(Equation 4)}$$

Where, φ_o is base value, φ_i represents SHAP value for feature i, z_i represents feature presence.

5.4.2 SHAP beeswarm plot

The SHAP beeswarm plot has features in the Y-axis and impact score in the X-axis. Impact score defines by what magnitude a feature pushes a prediction instance towards either positive or negative class. For each feature in the

Y-axis, the plot creates an array of red and blue dots spread across the impact score. The spread is intuitive. Blue dots represent lower magnitude while red dots represent higher magnitude. For integer features after scalar normalisation, values after 0 i.e (0,1] are represented by red dots whereas values before 0 i.e. [-1,0] are blue. Similarly, for categorical, encoded features, 0 (False) is blue while 1 (True) is red. For example, consider htn_no (No Hypertension) if the value is 1 (True), meaning the patient doesn't have hypertension, it will show as a red dot in the beeswarm plot for this particular instance (row), however the dot most probably will lie in the left side of the plot as having no hypertension actually reduces the chances of diseases. The main gist is that the colors of the dots represent that feature's magnitude for a particular instance and the position of the dot in the horizontal axis represents whether the feature pushes a prediction towards the positive class (ckd) or negative class (not ckd).

II. Results and Discussion

6.1. Performance Parameters

After all the preprocessing step and training of the model using Extra Trees and LightGBM, following parameters were obtained:

Table 3.

Cross-Validation (5-fold) Performance Parameters

Parameter	Extra Trees	LightGBM
Accuracy	0.993 ± 0.010	0.985 ± 0.018
Precision	1.000 ± 0.000	0.984 ± 0.033
Recall	0.988 ± 0.016	0.992 ± 0.010
F1 Score	0.994 ± 0.008	0.988 ± 0.016

6.1.1. Confusion Matrices

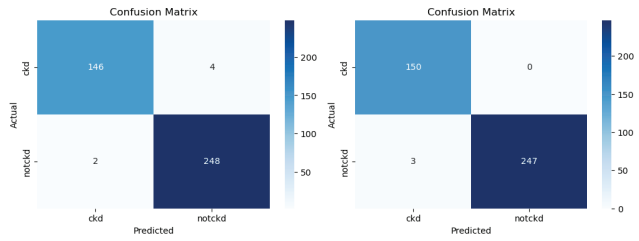

Figure 3. Confusion Matrix for (a) Extra Trees Model and (b) LightGBM

6.1.2. SHAP Visualisations

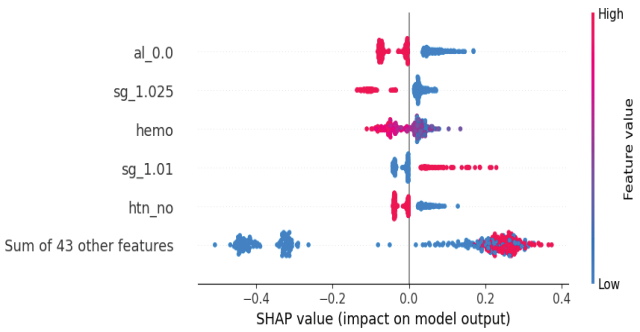

Figure 4. Global SHAP for Extra Trees Model

Figure 5. Global SHAP for LightGBM Model

6.2. Discussions

All performance parameters were optimal which means the model could accurately classify whether a person has CKD or not based on his blood test. Furthermore the SHAP visualisation helped us understand which features impacted the result of the prediction i.e whether a person has CKD or not, the most.

In the SHAP beeswarm plot, Blue dots represent lower magnitude while red dots represent higher magnitude, whereas the position of the dot in the horizontal axis represents the impact of the dot (feature in 'x' instance) towards the model prediction.

From *Fig. 4*, it can be seen that al_0.0 (Albumin=0.0) is the feature with highest importance. In this feature row, blue dots i.e 0 (False), are spread after 0.0 (base value) which means the patient having more than 0 albumin in the urine

test have higher chance of "ckd"while red ones i,e 1 (True) meaning, the patient has albumin=0.0 in the urine test, so he has higher likelihood of "notckd". This is inline with the current medical analysis and analytically shows that kidney not filtering properly which releases creatinine into the urine stream can be considered a predictor of kidney problems, possibly initial stages of CKD.

Clinically, Kidney produces erythropoietin, responsible for stimulating red blood cell production, so impaired kidneys reduces amount of hemoglobin in the blood. Fig. **5,** it can be seen that the model shows that having high hemo (red dots) means less chance of CKD while low hemo (blue dots) means high chance of CKD, this shows the model's ability to be used inversely to identify hidden symptoms and relation factor as a diagnostic tool, such as kidney impairment and reduced hemo can be connected by reduction in production of erythropoietin which may possible be due to kidney diseases.

In *Fig. 4.* hemo (Hemoglobin), al_0.0 (albumin=0) and sg_1.025 (specific gravity=1.025) served as the three highest impactful features. From fig **5,** al (Albumin), hemo (Hemoglobin) and sc (Serum creatinine) were the highest three. Having hypertension, less hemoglobin and more serum creatinine in the urine increasing the likelihood of CKD is logical and consistent with the medical theory. Similarly in fig **5, (**dm_no**)** pushes the prediction towards CKD when 0 (False and Blue Colored) i.e. a patient has diabetes, while pushes towards no CKD when 1 (True and Red) i.e.patient doesn't have diabetes. Other features such as various specific gravity values such as 1.025 and 1.015 can be considered unusual features considered important by the algorithm offering some significant medical insights.

III. Conclusion and Future Works

7.1. Conclusion

This study puts forwards a computationally efficient and lightweight framework integrable clinically with blood and urine tests using Extra Trees and LightGBM for early CKD detection, with high accuracies of 0.993 and 0.985 respectively using 5-fold Cross Validation. SHAP visualisation (Fig 4 and Fig 5) identified albumin, hemoglobin, specific gravity (1.025), hypertension, and serum creatinine as key predictors.

Low hemoglobin—with linkage to reduced erythropoietin production due to impaired kidneys—emerged as a critical biomarker, displaying this model's unique ability to uncover hidden clinical relationships. This SHAP-driven method can be used as a novel method for identifying other biomarkers, reproducible with other diseases as well. Unlike broader studies [7], this framework and reproducible UCI dataset enhance medical uses for routine CKD screening. The novel insight into specific gravity (1.025) suggests new medical research avenues as well.

7.2. Future Works

- Optimize hyperparameters to increase its performance on the dataset.
- Collect and validate for larger and more diverse CKD datasets
- Collaborate with clinical experts to investigate the medical significance of novel predictors like specific gravity.
- Develop a framework for deploying the model in real-time clinical systems.
- Apply similar SHAP insights for other diseases to identify hidden causes.

Acknowledgements

I would like to sincerely thank the following people and institutions for their assistance during this study:

- **Supervisor**: I am extremely grateful to Rajad Shakya sir for all of his help, advice, and criticism during this research. His knowledge and unwavering support were crucial in determining the course of this study.
- **The UCI Machine Learning Repository:** For the Chronic Kidney Disease dataset, which was crucial to this analysis, I am grateful to the UCI Machine Learning Repository.

References

[1] Poudel, S., 2025. *Prediction of chronic kidney disease using Random Forest, XGBoost and ANN model*. *Journal of Advanced College of Engineering and Management*, 10(1), pp.121–133. Available at: https://doi.org/10.3126/jacem.v10i1.76323 [Accessed 23 Jul. 2025].

[2] Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning*, 63(1), pp.3–42.

[3] Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32.

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, pp.3146–3154.

[5] Lundberg, S.M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp.4765–4774.

[6] Rubini, L., Soundarapandian, P. and Eswaran, P., 2015. *Chronic Kidney Disease [dataset]*. UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C5G020 [Accessed 23 Jul. 2025].

[7] Ghosh, S.K. and Khandoker, A.H., 2024. Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, 14(1), p.3687. https://doi.org/10.1038/s41598-024-54375-4