Chronic Kidney Disease Prediction

A *Machine Learning Approach Using NHANES Data*

Yung-Shan Chen

March 2025

Abstract

This capstone project investigates the use of supervised machine learning techniques to predict chronic kidney disease (CKD) using structured clinical data. Motivated by a personal experience and the clinical importance of early detection, the study leverages data from the National Health and Nutrition Examination Survey (NHANES), combining three cycles (2013–2020) with 24,132 samples. To improve sensitivity in identifying early-stage CKD, the project adopts a threshold of ACR $>$ 20 mg/g, as supported by clinical literature, and simulates real-world scenarios where urine samples are not always available by relying exclusively on blood test and demographic variables.

Five classification models—Logistic Regression, Random Forest, XGBoost, Support Vector Machine (RBF), and Deep Neural Network (DNN)—are implemented and evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC. Among them, XGBoost achieves the highest ROC-AUC (0.733), while the upgraded DNN model yields the highest recall (0.83), highlighting the trade-off between interpretability and sensitivity. SHAP analysis further reveals clinically meaningful predictors such as hemoglobin, creatinine, and blood urea nitrogen.

Despite promising results, the current models remain insufficient for direct clinical use. This study lays the groundwork for future multimodal approaches incorporating imaging and automated machine learning (AutoML) to improve performance and clinical applicability. The project not only demonstrates technical feasibility but also emphasizes the value of interpretable AI in healthcare.

1 Introduction

Chronic kidney disease (CKD) is a progressive condition that affects hundreds of millions of people worldwide. In its early stages, CKD is often asymptomatic, and without timely detection and intervention, it can lead to irreversible kidney damage, requiring dialysis or transplantation. Therefore, **early detection and risk prediction** play a crucial role in slowing disease progression and improving patient outcomes.

This project was inspired by a personal experience. In late 2024, I lost my beloved cat to kidney failure. The disease went unnoticed until it was too late for effective intervention, leaving a profound impact on me. This experience motivated me to explore how data science

and machine learning could be leveraged to assist in the early detection of kidney disease and prevent similar situations in the future.

As the capstone project for the IBM Supervised Machine Learning course, this study aims to develop a classification model to predict CKD risk based on clinical data. In addition to optimizing predictive performance, the project places strong emphasis on **model interpretability**, which is particularly important in healthcare applications where transparency and trust are essential.

This report details the entire modeling process, including data preprocessing, feature engineering, model selection, and performance evaluation. Furthermore, SHAP (SHapley Additive exPlanations) is employed to analyze feature contributions, ensuring that the model's predictions align with clinically meaningful factors.

Although the dataset used in this study is relatively limited in size, the project serves as a prototype for future multimodal learning approaches that integrate structured data with additional modalities, such as ultrasound imaging and electronic health records. The ultimate goal is to provide a potential predictive framework for settings where urine samples are difficult to obtain, such as in veterinary medicine or resource-limited regions.

2 Data Sources and Preprocessing

This study utilizes data from the National Health and Nutrition Examination Survey (NHANES), combining three cycles: 2013–2014, 2015–2016, and 2017–2020, with a total of 24,132 samples covering blood tests, urine tests, and demographic variables. Due to missing values for the gold standard diagnostic marker for kidney disease, the Albumin-to-Creatinine Ratio (ACR), in certain NHANES cycles, this study leveraged the naming conventions of NHANES data files to successfully retrieve unpublished urine data (e.g., DEMO_J.XPT, P_ALB_CR.XPT), ensuring the completeness of label definitions and enhancing the reliability and representativeness of the training dataset.

CKD Label Definition: According to standard clinical guidelines, Chronic Kidney Disease (CKD) is defined as ACR $>$ 30 mg/g. However, to facilitate the early detection of kidney damage, this study adopts a lower threshold of ACR $>$ 20 mg/g, based on recommendations from the literature[1]. According to the KDOQI guidelines and related research, an ACR in the range of 10–30 mg/g may already indicate early kidney impairment, particularly in high-risk populations such as those with diabetes or hypertension, who require closer monitoring[2]. Therefore, this study defines CKD as ACR $>$ 20 mg/g to enable earlier identification of at-risk individuals and promote early intervention and treatment.

Data Processing Workflow:

1. Merged data from the three NHANES cycles, standardizing column names and formats.
2. Standardized the label definition (using ACR $>$ 20 mg/g as the CKD criterion) and removed samples with missing labels.

[1]National Kidney Foundation. (2002). KDOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases*, 39(2 Suppl 1), S1–S266.

[2]Kramer HJ, et al. (2003). Increasing prevalence of chronic kidney disease in the United States. *Kidney International*, 64(4), 1409–1414.

3. Missing value imputation: Numerical variables (e.g., blood glucose, creatinine) were imputed using median values, while categorical variables (e.g., ethnicity, gender) treated missing values as a separate category. The overall missing data rate was below 10%, mainly concentrated in variables like red blood cell count and blood glucose.

4. Categorical variables were encoded using one-hot encoding to retain category information.

5. Numerical variables were standardized using StandardScaler to ensure stable performance in models sensitive to feature scaling, such as Logistic Regression and Deep Neural Networks (DNN).

6. Class imbalance adjustment: The merged NHANES dataset exhibited a low proportion of CKD-labeled (class 1) samples. Without adjusting the class distribution, the model would be biased toward predicting the majority non-CKD class (class 0), potentially leading to high false-negative rates and failing to detect CKD early. Therefore, under-sampling was performed to adjust the number of non-CKD samples to 1.5 times the CKD samples, resulting in an approximate 1.5:1 ratio. To address class imbalance, under-sampling was applied to the majority class. After this process, the dataset included 10,775 samples, comprising 4,310 CKD and 6,465 non-CKD cases. This adjustment helps improve the model's ability to identify CKD cases, particularly enhancing critical metrics such as recall and ROC-AUC.

Feature Selection and Data Quality: To prevent the model from overly relying on urine biomarkers (e.g., ACR, UMA, UCR), this study retained only blood test and demographic variables as features, simulating clinical scenarios where urine samples may not be available, thereby broadening the model's applicability. This approach is supported by literature[3], demonstrating that even without urine samples, effective prediction of CKD progression can be achieved using demographic and laboratory data alone, enhancing the model's clinical utility.

Finally, the dataset was split into 80% training and 20% testing sets, stratifying by the CKD label to ensure that the class distribution remains consistent across both sets, thus maintaining stability and reproducibility in model training and evaluation.

3 Modeling and Evaluation

To evaluate the predictive performance for chronic kidney disease (CKD) detection, this study implemented and compared five supervised learning models that represent a spectrum of interpretability and modeling capacity:

- **Logistic Regression**: as a baseline model with high interpretability.

[3]For example: Wang F, et al. (2019). Predicting kidney function decline in patients with CKD using only demographic and laboratory data.
Journal of the American Society of Nephrology, 30(10), 1961–1970.

- **Random Forest**: to capture non-linear relationships and interactions between features.
- **XGBoost**: an ensemble gradient boosting method known for its robust predictive performance.
- **Support Vector Machine (SVM) with RBF kernel**: to enhance classification boundaries in non-linear space.
- **Deep Neural Network (DNN)**: to explore high-dimensional non-linear patterns.

3.1 Handling Class Imbalance

The dataset exhibited class imbalance, with significantly more negative (non-CKD) samples than positive (CKD) samples. To mitigate bias in the model toward the majority class, different techniques were applied across models:

- **Logistic Regression**, **Random Forest**, and **SVM**: incorporated `class_weight` adjustments ({0: 0.8, 1: 1.2}).
- **XGBoost**: employed `scale_pos_weight` based on the negative-to-positive sample ratio to re-balance the objective function.
- **DNN**: utilized **Focal Loss** (gamma: 2, alpha: 0.5) to focus learning on hard-to-classify minority samples.

3.2 Model Configurations and Hyperparameters

Logistic Regression:

- Regularization: L2 penalty
- Regularization strength: 5.0
- Solver: liblinear

Random Forest:

- Number of estimators: 300
- Maximum depth: 8
- Minimum samples per split: 20
- Minimum samples per leaf: 10

XGBoost:

- Number of estimators: 400
- Maximum depth: 4
- Learning rate: 0.03
- Subsample ratio: 0.8
- Column sampling by tree: 0.8

SVM (RBF Kernel):

- Regularization parameter (C): 1.0
- Kernel coefficient (γ): `scale` (automatically computed)
- Probability estimation: enabled for AUC calculation

Deep Neural Network (Upgraded Version):

- Architecture: 4 hidden layers (256, 128, 64, 32 units respectively)
- Activation function: Leaky ReLU in hidden layers
- Regularization: L2 penalty (λ: 0.001)
- Dropout rate: 0.1
- Optimizer: RMSprop (learning rate: 0.0005)
- Loss function: Focal Loss (gamma: 2, alpha: 0.5)
- Learning rate scheduler: ReduceLROnPlateau
- Early stopping: patience: 10
- Classification threshold: 0.42

3.3 Data Splitting and Evaluation Metrics

The dataset was split into 80% training and 20% testing subsets, maintaining the class distribution using stratified sampling. The following evaluation metrics were employed to assess model performance:

- **Accuracy**: the overall proportion of correct predictions.
- **Precision**: the proportion of true positives among predicted positives.
- **Recall**: the proportion of true positives identified among actual positives.

- **F1-score**: the harmonic mean of precision and recall.
- **ROC-AUC**: the area under the Receiver Operating Characteristic curve, measuring the model's ability to discriminate between classes across different thresholds.

Given the clinical importance of minimizing false negatives in CKD detection, emphasis was placed on **recall** and **ROC-AUC** as key indicators of model performance.

4 Results and Comparison

To evaluate the effectiveness of different supervised learning models for CKD prediction, each model was trained and tested on the same dataset splits, and their performance metrics were recorded. The following metrics were used: Accuracy, Precision, Recall, F1-score, and ROC-AUC.

4.1 Performance Comparison

Table 1: Performance Metrics for CKD Prediction Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Logistic Regression	0.66	0.57	0.61	0.59	0.709
Random Forest	0.68	0.62	0.54	0.58	0.726
XGBoost	0.68	0.59	0.61	0.60	0.733
SVM (RBF Kernel)	0.68	0.59	0.62	0.60	0.724
DNN (Upgraded)	0.61	0.51	**0.83**	0.63	0.724

4.2 Analysis of Results

The **XGBoost** model demonstrated the highest **ROC-AUC** score of 0.733, indicating superior discriminative ability across thresholds. Both **XGBoost** and **SVM** achieved balanced performance in terms of precision and recall.

The **DNN** model, although having a lower accuracy (0.61) and ROC-AUC (0.724), achieved the highest **recall** (0.83), making it potentially useful for reducing false negatives, which is critical in healthcare scenarios. However, the DNN's performance exhibited greater variability due to its sensitivity to random initialization.

In contrast, tree-based models like **Random Forest** and **XGBoost** provided more stable results with higher precision, making them suitable for settings where the cost of false positives must also be considered.

Considering both performance stability and interpretability, **XGBoost** was selected as the primary model for further interpretability analysis using SHAP, while the DNN results serve to highlight the potential benefits and trade-offs of deep learning in this context.

5 Interpretability with SHAP

Given the critical need for interpretability in healthcare-related applications, the **XGBoost** model was further analyzed using **SHAP (SHapley Additive exPlanations)**. SHAP assigns each feature an importance value for a particular prediction, providing insights into how individual features contribute to model outputs.

5.1 Feature Importance Summary

Figure 1 presents the SHAP summary plot, which displays the overall contribution of each feature across the dataset. Notably, the most influential features in predicting CKD included:

- **Hemoglobin (LBXHGB)**: lower levels are strongly associated with higher CKD risk.
- **Serum Creatinine (LBXSCR)**: elevated levels indicate impaired kidney function.
- **Blood Urea Nitrogen (LBXSBU)**: higher concentrations are linked to renal stress.
- **Demographic Variables (e.g., RIDRETH1 - Race/Ethnicity)**: highlight potential health disparities.

These findings align with known clinical indicators for kidney function, reinforcing the biological plausibility of the model's predictions.

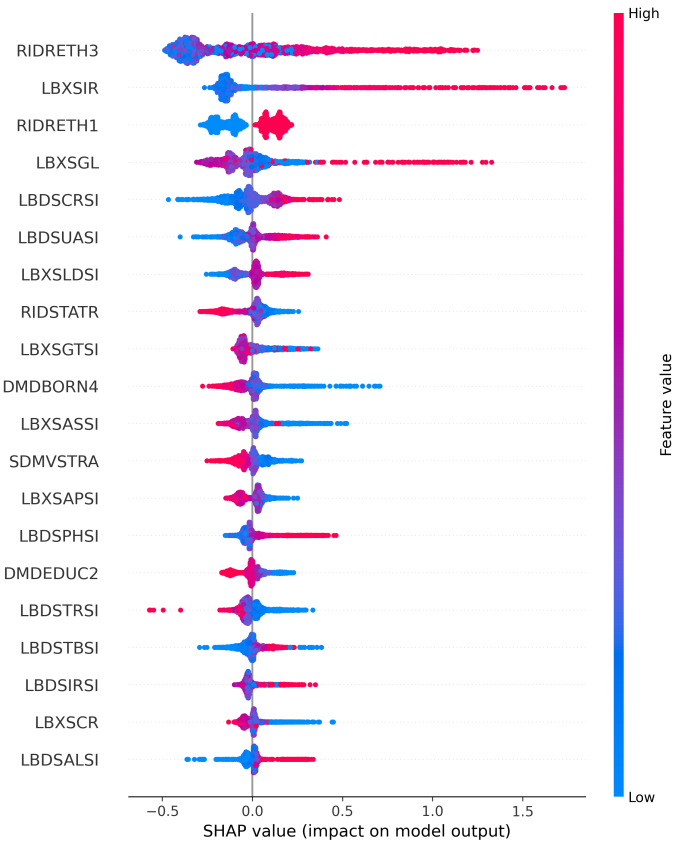

Figure 1: SHAP summary plot illustrating feature contributions to CKD predictions.

5.2 Feature Interaction: Dependence Plot

To explore feature interactions in more detail, a SHAP dependence plot was generated for **Blood Uric Acid (LBXSUA)**, which appeared among the top features. Figure 2 illustrates that uric acid levels interact with other variables, such as demographic factors, in influencing CKD risk predictions. This suggests that elevated uric acid alone is not uniformly predictive of CKD but may have varying effects depending on other patient characteristics.

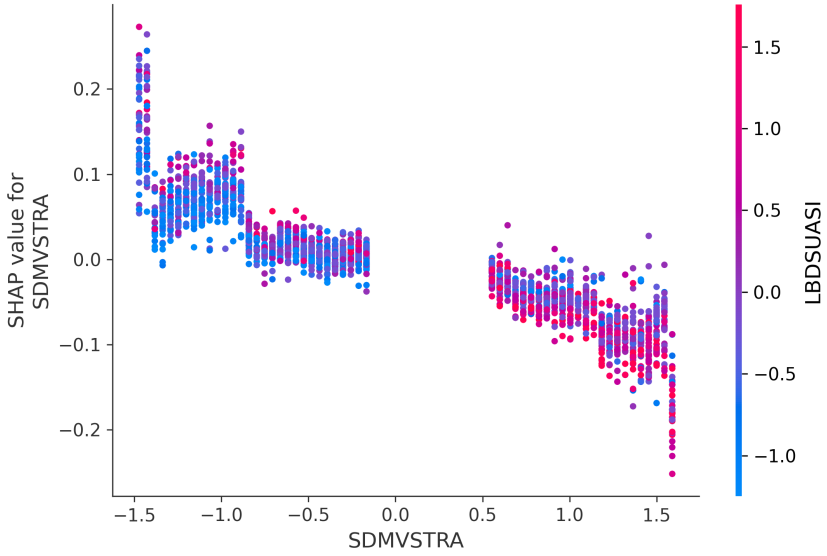

Figure 2: SHAP dependence plot for Blood Uric Acid (LBXSUA), showing interactions with other features.

5.3 Insights and Clinical Relevance

The SHAP analysis confirmed that the XGBoost model's decision-making process aligns with established medical knowledge. Key biomarkers related to kidney function, such as hemoglobin and creatinine, were identified as the most influential features, ensuring that the model's predictions are grounded in clinically meaningful variables.

This interpretability approach not only enhances trust in the model but also provides actionable insights for healthcare practitioners, potentially guiding further diagnostic assessments or interventions.

6 Key Findings and Insights

This study implemented and compared multiple supervised learning models for chronic kidney disease (CKD) prediction. The following key insights were derived from model performance and interpretability analyses:

6.1 Model Performance Summary

Among all models, **XGBoost** achieved the highest **ROC-AUC score (0.733)**, indicating strong and stable performance in distinguishing CKD from non-CKD cases. It also maintained a well-balanced trade-off between precision and recall, making it a suitable candidate for clinical screening.

The upgraded **DNN** model achieved the highest recall **(0.83)**, showing strong potential in identifying CKD-positive cases. However, this advantage came with notable variability across training runs, suggesting that the model's performance may be less consistent than more stable alternatives like XGBoost.

Random Forest and **SVM (RBF)** showed stable performance, with AUC scores between 0.72 and 0.73. These models demonstrated higher precision but relatively lower recall compared to DNN.

6.2 Feature Importance and SHAP Analysis

SHAP analysis revealed that the most influential features for CKD prediction included:

- **Hemoglobin (LBXHGB)**: Lower levels were strongly associated with increased CKD risk, consistent with anemia commonly observed in kidney disease.
- **Serum Creatinine (LBXSCR)**: Elevated levels served as a critical marker of renal impairment.
- **Blood Urea Nitrogen (LBXSBU)**: High levels indicated reduced kidney clearance.
- **Uric Acid (LBXSUA)**: Its impact varied depending on interactions with demographic variables such as race/ethnicity.
- **Race/Ethnicity (RIDRETH1)**: Disparities in CKD risk were observed across population groups, reflecting potential health inequities.

These findings are consistent with established clinical knowledge and demonstrate that the models successfully identified biologically meaningful predictors of CKD.

6.3 Clinical Implications

This study highlights the potential of machine learning to predict CKD using only blood tests and demographic features—particularly valuable in settings where urine test results are unavailable.

- For applications where reducing **false negatives** is a top priority, the **DNN** model may be preferred due to its high recall, making it suitable for **sensitive early screening**.
- For applications requiring **stability and interpretability**, **XGBoost** offers the best trade-off and can be paired with SHAP to support transparent decision-making in clinical settings.

7 Conclusion and Future Work

This study demonstrated that machine learning models can effectively predict chronic kidney disease (CKD) using structured clinical data, even in the absence of urine test results. Among the evaluated models, **XGBoost** provided the best balance between predictive performance and interpretability, while the **Deep Neural Network (DNN)** model achieved the highest recall, offering potential advantages for early screening by reducing false negatives.

However, from a clinical perspective, **the predictive accuracy of the current models remains insufficient for direct clinical deployment**. This outcome highlights two key limitations: first, the dataset primarily comprises health screening samples, which may not fully capture the diversity of CKD progression; second, the CKD labels were defined using the albumin-to-creatinine ratio (ACR), which, while widely accepted, has limitations in detecting early-stage kidney dysfunction. To enhance the clinical utility of such models, future work may need to adopt more accurate labeling methods, such as imaging-based diagnoses, to replace or supplement ACR.

Given the often asymptomatic nature of early CKD, regular health checkups are typically required for early detection. This project aimed to explore whether routine blood test variables could serve as an early warning system, helping physicians decide whether further diagnostic tests are warranted. This approach could also benefit veterinary applications, particularly for animals like cats and dogs where urine samples are difficult to obtain, making ACR measurements impractical. Nevertheless, **the current model's precision is not yet sufficient to support clinical decision-making independently**.

Future research directions include:

- **Multimodal learning**: Integrating ultrasound imaging with structured data to develop multimodal deep learning models that enhance CKD detection. As demonstrated in related studies (see Appendix), deep learning applied to ultrasound images has already achieved high diagnostic accuracy, suggesting a promising path for clinical applications.
- **Label optimization**: Utilizing imaging-based diagnoses as a replacement for ACR in defining CKD labels, to improve the clinical relevance of model predictions.
- **Automated machine learning (AutoML)**: Leveraging AutoML platforms such as Google AutoML or H2O for efficient model selection and hyperparameter tuning.

This project serves as an exploratory implementation, providing a prototype for machine learning-based CKD prediction. With access to larger multimodal datasets and more accurate labeling, future work could lead to the development of clinically viable predictive systems.

Appendix: Personal Motivation: A Tribute to My Cat

Last year, I lost my beloved cat to kidney failure. I failed to notice the subtle early signs, and by the time she was taken to the hospital, it was already too late. Despite our best efforts and aggressive treatment, we could not stop death from taking her away—just three years into her life.

This report is dedicated to her memory. One day, I hope to design a system capable of providing early warnings for kidney failure. Though I can no longer save her, perhaps this work may help save someone else's beloved companion.

Figure 3: My beloved cat who inspired this research project.