Derivation of a risk score arising from an analysis of potential risk factors for prediction and analysis of chronic kidney disease (CKD)

Bryce Parkman
University of Massachusetts Amherst
Amherst, MA, USA
bparkman@umass.edu

Manan Talwar
University of Massachusetts Amherst
Amherst, MA, USA
mtalwar@umass.edu

Abstract—*This paper attempts to describe the observations and results obtained from a supervised machine learning classification algorithm geared towards predicting chronic kidney disease using physiological factors. Our results establish that ACR and GFR continue to be the most prominent indicators of CKD, and that glycohemoglobin is a significant factor in the diagnosis of CKD. Our work also suggests that presence of Manganese in urine seems to have a pronounced significance in the diagnosis of CKD using our model as compared to other metallic elements considered in this study. The calculated risk score was based on these significant factors and could allow patients to better evaluate their CKD risk.*

Keyword—Chronic Kidney Disease (CKD), Supervised Machine Learning, NHANES, GFR, ACR

I. INTRODUCTION

Kidney diseases are a leading cause of death in the United States. [3] The CDC estimates that more than 1 in 7 people, accounting for approximately 15% of US adults or 37 million people have CKD. Medical literature has established that chronic kidney disease (CKD) is a multifactorial disease related to gender, age, obesity and smoking, hypertension and cardiovascular diseases and also to genetic and environmental factors. In recent time, extensive research has shown that Type-1 and Type-2 diabetes, high blood pressure, the presence of many different heavy metals either in blood or in urine and more recently even demographic factors such as race are known to be risk factors for acute as well as chronic kidney disease. While widely recognized, this risk has been studied in very limited and controlled settings. [3].

II. OBJECTIVES

The aim of this study is to quantify a patient's kidney disease risk based on a combination of physiological factors. These include, but are not limited to, heavy metal presence in blood and urine, urine albumin-to-creatinine ratio (uACR), glomerular filtration rate (GFR), blood sugar levels, and prior history of cardiovascular disease as input variables. We formulated the following research questions, which our study aims to address:

RQ-1: Which of the input variables under consideration can be considered significant in the diagnosis of CKD?

RQ-2: Does the presence of metallic elements in either blood or urine affect the diagnosis of CKD?

III. BACKGROUND

Albumin is a biological protein made by the liver that is found in human blood. It is important that it stays in the blood since it helps serveral molecules travel through the bloodstream and prevents fluid leakage. High levels of albumin in the urine is an indicator for kidney disease since the kidney is not meant to filter this into urine [3].

Creatinine is a biological protein created as a waste product by muscles that the kidney is meant to filter into the urine [3]. Hence, a low creatinine levelin urine is considered to be an indicator of kidney disease. It should be noted that there is a difference between looking at creatinine in the bloodstream (called "serum creatinine") and looking at creatinine in the urine (called "creatinine clearance") which are both obtained from two different lab tests.

The ratio between urine albumin and urine creatinine is known as uACR. Since both Albumin and Creatine are indicators of kidney disease, uACR is used as a marker for kidney health which combines the effect of both Creatinine and Albumin. The higher the ACR, the worse the kidney may be performing.

$$\frac{\text{Urine albumin}\left(\text{mg dL}^{-1}\right)}{\text{Urine creatinine}\left(\text{g dL}^{-1}\right)} = \text{UACR}\left(\text{mg g}^{-1}\right) \quad (1)$$

Glomelur Filtration Rate (GFR) is the rate at which the kidney filters waste from the bloodstream. In medical literature, GFR is the most reliable indicator of kidney health. A low GFR is almost always caused either by kidney disease or in more severe cases kidney failure [2]. However, this value is hard to calculate precisely since clinicians measure how well your kidneys are filtering certain agents not produced by your body, such as inulin (a kind of fiber that is found in some plant foods) and iohexol (contrast agent used in imaging tests) [2]. It is for this reason that healthcare professionals use a formula to estimate GFR (eGFR). To do so, a calculation is used to estimate how well the kidneys are filtering certain agents produced by the body, such as creatinine (a waste product that comes from the normal wear and tear on muscles) and cystatin C (a protein that slows down the breakdown of other protein cells) [2]. A commonly used method to estimate GFR is to use the CKD-EPI equation since it generally gives better results over other formulas. S_{cr} is serum creatinine (in blood), and k and α are values dependent on the patient's gender.

$$\text{GFR} = 141 \times \min(S_{Cr}/\kappa, 1)^{\alpha} \times \max(S_{Cr}/\kappa, 1)^{-1.209} \times 0.993^{Age} \times (1.018 \text{ if female}) \times (1.159 \text{ if African American}) \quad (2)$$

On their own, ACR and GFR can only give insight into kidney health when they are at extreme values. However, taking them in conjunction can give a much bigger picture. According to the National Kidney Foundation, the criteria for a clinical diagnosis of CKD is defined as under:

Criteria for CKD: Either of the following present for > 3 months	
Markers of kidney damage (one or more)	• Albuminuria (ACR ≥ 30 mg/g) • Urine sediment abnormalities • Electrolyte and other abnormalities due to tubular disorders • Abnormalities detected by histology • Structural abnormalities detected by imaging • History of kidney transplantation
Decreased GFR	GFR <60 ml/min/1.73 m^2

Table 1: CKD clinical diagnosis criteria

IV. METHODS AND OBSERVATIONS

A. Data collection

Our source for data in this study is the NHANES [1] dataset, specifically the continuous NHANES datasets from 2009 - March 2020 (pre-pandemic). We derived our input variables as follows. ACR (mg/g) is collected directly from the dataset, whereas GFR (ml/min/1.73m^2) is estimated using the CKD-EPI equation [2] which factors in age, gender, race, and blood serum creatinine level. It may be noted that creatinine data is only available for correspondents >=20 years of age, so our analysis is limited to correspondents who fit this criteria.

Other input variables such as Hb1ac, cholesterol, blood pressure, and presence of heavy metals in urine (barium, manganese, cadmium, cobalt, caesium, lead, tin, and nolybdenum) were also directly sourced from the NHANES dataset. The questionnaire data includes the "Kidney Conditions Questionnaire" dataset. We were only interested in the variable KIQ022 ("Have you ever been told by a healthcare professional that you had weak or failing kidneys?"). We did not use this parameter as the target, rather as an input variable.

B. Indicator and Target Variable

Based on our analysis, we created an indicator for identifying patients having kidney disease based on the diagnosis criteria for CKD [3] as under:

- ACR ≥ 30 mg/g
- GFR < 60 ml/min/1.73m^2.

We applied this indicator on our dataset to generate the 'TARGET' label. After applying this indicator, we noticed the dataset was very skewed with only 3.75% of the subjects classified as having CKD.

C. Dataset Preprocessing and Preparation

We manually screened the dataset for potential vulnerabilities that could cause errors in our model. We performed the following cleaning measures:

- We removed all duplicate instances that existed in the dataset.
- We removed all instances which had at least one 'None', 'Nan' or 'null' value.
- We identified categorical variables and converted them to numerical attributes using one hot encoding ('KIQ022, RIDRETH1, RIAGENDR).
- We normalized the numerical attributes in the dataset using the MinMaxScaler

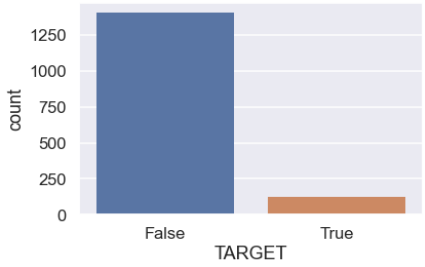

Figure 1: Imbalance in dataset

We identified that our dataset was extremely unbalanced and such a dataset could impact the model performance negatively. Therefore, we decided to adjust this imbalance by using the state-of-the-art SMOTE technique by oversampling the minority class (labeled as True in the 'TARGET' attribute). This eradicated the imbalance and our dataset was prepared for further analysis.

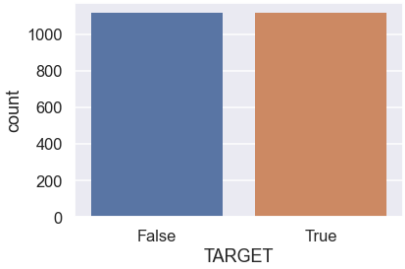

Figure 2: Dataset after SMOTE

D. Feature Selection

Dimensionality reduction was necessary for our study since we were dealing with a high-dimensional dataset which could have led to overfitting. Since the relationships between the various attributes are complex, we decided to use a combination of feature selection techniques:

- We used the chi-squared test to quantify the relationship between two attributes. We then used ANOVA-F test to determine whether the variability between group means is larger than the variability of the observations within the groups.
- We calculated the mutual information, which is the "amount of information" obtained about one random variable by observing the other random variable.

We identified overlapping features from the three tests above to select features for our model:

LBXSIR, LBXTC, URXUBA, URXUMN, LBXGH, LBXIN, URXUMA, URXUCR, RIDAGEYR, RIDRETH1_1.0, LBXSTP, LBXSCR, RIAGENDR_2.0

These features correspond to the following: iron in blood ,total cholesterol, barium in urine, manganese in urine, glycohemoglobin, insulin, albumin in urine, creatinine in urine, age, race as Mexican American, total protein in blood serum, and gender as female.

E. Model Evaluation

We screened several machine learning models as potential candidates. However, keeping in mind that we needed both a label as well as a numeric score, we decided to consider Logistic Regression and Random Forest since the probability for each label is easily accessible for both these algorithms using sklearn's predict_proba function[6]. The probability for a positive diagnosis of CKD will be our risk score for the patient. We split the processed dataset into two parts: 80% and 20% the original size for training and testing respectively. Using the training data, we created an ensemble of both the algorithms and evaluated them using Stratified K-Fold Cross Validation Technique.

	Random Forest	**Logistic Regression**
Accuracy	0.986	0.688
Precision	0.980	0.689
Recall	0.994	0.684
F-1	0.987	0.684
ROC-AUC	0.999	0.688

Table 2: Observed metrics during CV

Figure 3: F-1 Score variation over CV runs

V. RESULTS

Based on our observations above, we selected the Random Forest classification model and obtained the following classification report:

	precision	recall	f1-score	support
False	1.00	0.99	0.99	286
True	0.90	0.95	0.93	20
accuracy			0.99	306
macro avg	0.95	0.97	0.96	306
weighted avg	0.99	0.99	0.99	306

Figure-4: Classification Report for RF

Feature Significance was obtained as under:

```
[0.03976047 0.04467963 0.04900621 0.03068332 0.06484812 0.05586978
 0.38433634 0.11140765 0.05322152 0.0049665  0.04608897 0.10348815
 0.01164333]
```

Figure - 5: Feature Importance

Using the aforementioned predict_proba() function, we were able to get probability scores for CKD diagnosis being true/false. This is depicted below for a small sample of patients, where the left value is for positive diagnosis and the right value for negative. The probability for positive diagnosis is our risk score.

```
[[0.95 0.05]
 [0.92 0.08]
 [0.99 0.01]
 [1.   0.  ]
 [1.   0.  ]
```

Figure - 6: A subset of Probability Scores

VI. DISCUSSION

For our first research question, we observed feature importances obtained from our model. Results suggest that albumin and creatinine in urine emerge as decisive factors in the diagnosis of CKD. They are followed by total protein in blood serum and glycohaemoglobin. Since prior research has established a relationship between diabetes mellitus and glycohemoglobin, we propose that future work should consider the relationship between diabetes, glycohemoglobin and CKD.

For our second research question, we explored the feature importances for attributes pertaining to metals being detected either in the blood serum or urine samples. Our study does not yield conclusive results about the role of the presence of metals in blood and/or urine on a diagnosis on CKD. Manganese in urine seems to have a more pronounced significance than any other detected metallic component, but future work should quantitatively assess such correlations.

Our risk scores were established on the basis of the probability of a particular label being labeled as CKD-Positive (True). We propose that future work should consider developing a more holistic measure that incorporates statistical significance of each significant attribute.

It is important to note that our study aimed at establishing merely quantitative relationships between a diagnosis of CKD and the attributes resulting from feature selection. Future work should consider exploring quantitative relationships including establishing statistical significance of relevant attributes through statistical hypothesis testing and other methods. It may also be noted that our results are based on a somewhat synthetic dataset due to oversampling which could be a threat to the validity of our claims.

VII. Conclusion

CKD is a major health risk that usually exhibits no symptoms until CKD is at an advanced stage[5], so it is important for people to know their risk before it is too late. We sought out to develop a risk score based on many physiological features so medical professionals and patients can get a better grasp of their kidney health. After determining which features were most important through multiple techniques, we were able to use a random forest model to create a risk score for each patient. While we identified a few shortcomings with our process, we still believe this risk score can be extremely useful for the medical community and hope future studies can improve upon our approach.

VIII. REFERENCES

[1] www.cdc.gov. 2022. *NHANES Questionnaires, Datasets, and Related Documentation*. [online] Available at: <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020> [Accessed 25 March 2022].

[2] Rate, E. and Health, N., 2022. *Estimating Glomerular Filtration Rate | NIDDK*. [online] National Institute of Diabetes and Digestive and Kidney Diseases. Available at: <https://www.niddk.nih.gov/health-information/professionals/clinical-tools-patient-management/kidney-disease/laboratory-evaluation/glomerular-filtration-rate/estimating> [Accessed 25 March 2022].

[3] Tsai, H., Hung, C., Wang, C., Tu, H., Li, C., Tsai, C., Lin, W., Chen, S. and Kuo, C., 2022. *Associations among Heavy Metals and Proteinuria and Chronic Kidney Disease*.

[4] ucsfhealth.org. 2022. *Lead Levels Blood*. [online] Available at: <https://www.ucsfhealth.org/medical-tests/lead-levels---blood> [Accessed 1 April 2022].

[5] www.cdc.gov. 2020. *What You Should Know About Chronic Kidney Disease.* [online] Available at: <https://www.cdc.gov/kidneydisease/publications-resources/what-to-know-about-ckd.html>. [Accessed 2 April 2022].

[6] scikit-learn.org. 2022. *sklearn.ensemble.RandomForestClassifier* [onlne] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.predict_proba> [Accessed 2 April 2022].