Samantha Zocher

Chronic Kidney Disease Prediction in the All of Us Dataset

HIDS 7006

**Introduction**

It is estimated that 35.5 Americans have Chronic Kidney Disease, or CKD1. Chronic Kidney Disease occurs when there is significant kidney damage, causing waste to build in the body. However, as many as 20 million of those people do not know they have it. CKD has 5 stages, each one causing more and more damage. Stages 1 and 2 typically have no noticeable symptoms, leading to late detection of CKD in many patients. Late detection has many negative effects, including a worse prognosis and irreversible effects. While CKD is not curable, early detection can significantly slow the progression of the disease and afford patients a higher quality of life. Once a patient notices symptoms and goes to the doctor, the condition is diagnosed by measuring glomerular filtration rate. If a patient has a glomerular filtration rate of <60 for at least 2 months, they would be diagnosed with CKD. In order to catch more cases early, machine learning studies have been conducted for this classification task. However, many of these studies involve small sample sizes and patients recruited from one hospital or geographic area. Using the All of Us dataset, a diverse, openly accessible health database containing EHR, survey, and genomic data2, a broadly applicable CKD classification model will be constructed.

**Literature Review**

A study conducted by Islam et al3 achieved a F1 score of 0.98 using 24 features and a XGBoost model. However, only 400 cases are included and the classes are balanced. This may not translate well into a clinical environment as there will not be an equal distribution of CKD and nonCKD patients in a real setting. Another study by Bahrami et al4 reported a F1 score of 0.94 using a stratified k-fold neural network. Fewer features were used in this model but there was a larger cohort size of 6,855. There was a class imbalance in this study, which replicates reality better. Debal & Sitote5 researched and implemented a similar random forest model with a F1 score of 0.78. With less than 2,000 patients from the same hospital, 19 features, and a goal of predicting CKD classes, a lower F1 score is to be expected. Finally, Bai et al6 conducted a retrospective cohort study on patients for their likelihood to develop end stage renal disease after 5 years of CKD progression. While this study had a slightly different objective than mine, the features they selected aided my own feature selection. All of these studies utilized laboratory and demographic variables. Some notable important features for these studies include hemoglobin, albumin, specific gravity, BMI, presence of hypertension, glucose, creatinine, and blood urea nitrogen. These previous works will inform what features are included in the model, what metric is optimized , and which models are tested.

**Methods**

Using the Controlled Access Tier for the All of Us dataset, the participant cohort was constructed. Participants were excluded from the study if they had cancer, due to outlier lab values that could affect the model’s performance. 2 classes, CKD and non-CKD, were constructed using EHR and lab data. If a patient had an applicable ICD-10 code for CKD or had a glomerular filtration rate <60 recorded at least twice, they were classified as CKD patients. 127,487 patients were included in the final analysis, with 24,166 patients having CKD and 103,321 patients not having CKD. This imbalance will be addressed later in the analysis.

Next, feature engineering and preprocessing was conducted. A total of 30 features were created, including different demographic, SDOH, genetic, and lifestyle factors. The complete list of features include ethnicity, race, sex, age, smoking status, alcohol score, physical health score, mental health score, BMI, diastolic blood pressure, systolic blood pressure, albumin, creatinine, glucose, potassium, sodium, blood urea nitrogen, RBC, hemoglobin, WBC, Alkaline phosphatase, calcium, chloride, proteinuria, anemia, diabetes, hypertension, coronary artery disease, whether a patient can afford a PCP, and polygenic risk score(PRS). The PRS was constructed from a previously constructed GWAS by Wutte et al7.

Continuous features were accessed for normality and transformed if they were non normal. Additionally, categorical variables were one-hot encoded or transformed to numerical mappings if they were ordinal. Finally, a standard scalar function was applied to all continuous variables. 



After all the features were processed, the outcome and feature set were split. After that, the data frames were split 80/20 into training and testing data. 5 different models were fitted to this data including a decision tree, random forest, XGBoost, AdaBoost, and a neural network.

**Results**

Without any hyperparameter tuning or feature selection, the XGBoost model performed the best, with an F1 score of 0.65. However, the most important evaluation metric for this problem was recall, as the goal was to capture any positive CKD patients. Precision was sacrificed to optimize recall, but it would be better to test more patients that ended up being flagged as false positives than to miss a CKD case. Using gridsearch, applicable hyperparameter tuning was performed. The best recall performing model had 25 features, a recall of 0.82, and a F1 score of 0.64. The confusion matrix in Figure 4 reveals that 4,000 CKD patients were correctly identified and only 892 CKD patients were missed by the model. While there is clear room for improvement, the recall score is much better than some previously conducted works. 



**Discussion**

Early prediction is extremely important to improve CKD awareness and long-term outcomes. Machine learning models like the one built in this study will help patients be aware of their condition before it progresses to a late stage. However, there are certainly limitations that come with the All of Us dataset. While the goal of All of Us is to create a more diverse study population, the majority of participants included in this model were still white. This may create a bias within the model where it performs well for white participants but poorly for non-white participants. Machine learning biases are important to catch and correct as possible. Once more participants are recruited to join All of Us, it would be useful to redo this study with a more diverse group of participants. While this study’s machine learning model results may not be robust enough for clinical implementation, it demonstrates a promising start for larger scale, generalizable classification models for CKD.

**Conclusion**

Chronic Kidney Disease prevalence is increasing in the United States with very little awareness of the condition in early stages. In order to improve patient outcomes, reduce healthcare costs, and slow progression, early detection is vital. This study aimed to create early detection methods through machine learning. A XGBoost model performed the best, with an accuracy of 0.82, precision of 0.52, recall of 0.82 and F1 score of 0.64. These findings were made possible by the participants enrolled in the All of Us study funded by the NIH, and as more participants join this study, this model could improve and represent a more diverse population.

**References**

1. Centers for Disease Control and Prevention. Chronic Kidney Disease in the United States, 2023. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2023.
2. All of Us Research Program Investigators; Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E. The "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668-676. doi: 10.1056/NEJMsr1809937. PMID: 31412182; PMCID: PMC8291101.
3. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inform. 2023 Jan 12;14:100189. doi: 10.1016/j.jpi.2023.100189. PMID: 36714452; PMCID: PMC9874070.
4. Bahrami, P., Tanbakuchi, D., Afzalaghaee, M. et al. Development of risk models for early detection and prediction of chronic kidney disease in clinical settings. Sci Rep 14, 32136 (2024). https://doi.org/10.1038/s41598-024-83973-5
5. Debal, D.A., Sitote, T.M. Chronic kidney disease prediction using machine learning techniques. J Big Data 9, 109 (2022). https://doi.org/10.1186/s40537-022-00657-5
6. Bai, Q., Su, C., Tang, W. et al. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep 12, 8377 (2022). <https://doi.org/10.1038/s41598-022-12316-z>
7. Wuttke, M., Li, Y., Li, M., Sieber, K. B., Feitosa, M. F., Gorski, M., Tin, A., Wang, L., Chu, A. Y., Hoppmann, A., Kirsten, H., Giri, A., Chai, J. F., Sveinbjornsson, G., Tayo, B. O., Nutile, T., Fuchsberger, C., Marten, J., Cocca, M., Ghasemi, S., … Pattaro, C. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics*, *51*(6), 957–972. https://doi.org/10.1038/s41588-019-0407-x