Prediction of Chronic Kidney Disease Degeneration with Machine Learning

Jamiree Harrison†[1], Manuchehr Aminian[2], Anna K. Berryman[3], Georgia S. Brennan[3], Zhaoshu Cao[4], Claire S. Chang[5], Quindel Jones[6], Yena Kim[7], Kimberly Matsuda[8], Brady Metherall[3], Sandra A. Tsiorintsoa[9], Nilofar Varzgani[10], Per Wagenius[11],

[1] *Department of Mechanical Engineering, the University of California Santa Barbara, USA*
[2] *Department of Mathematics and Statistics, California State Polytechnic University Pomona, USA*
[3] *Mathematical Institute, University of Oxford, UK*
[4] *Department of Mathematics, New Jersey Insititute of Technology, USA*
[5] *School of Operations Research & Information Engineering, Cornell University, USA*
[6] *Department of Mathematical Sciences, Virginia CommonWealth University, USA*
[7] *Department of Mathematics, Hawaii Pacific University, USA*
[8] *Department of Mathematical Sciences, Rensselaer Polytechnic Institute, USA*
[9] *Department of Mathematics and Statistics, Clemson University, USA*
[10] *Department of Business Systems & Analytics, La Salle University, USA*
[11] *Department of Mathematical Sciences, University of Vermont, USA*

(*Communicated to* MIIR *on 1 April 2024*)

Study Group: 39th Annual Workshop on Mathematical Problems in Industry, New Jersey Institute of Technology, 12–16 June 2023.

Communicated by: Linda Cummings, New Jersey Institute of Technology

Industrial Partner: Vironix Health

Presenter: Botros Toro

Team Members: Olayemi Adeyemi; Manuchehr Aminian; Anna K. Berryman; Georgia S. Brennan; Zhaoshu Cao; Claire S. Chang; Evelina Dubovski; Kossi Edoh; Jamiree Harrison; Olaniyi Iyiola; Quindel Jones; Yena Kim; Amy Lu; Kimberly Matsuda; Brady Metherall; Olaoluwa Ogunleye; Sandra A. Tsiorintsoa; Nilofar Varzgani; Per Wagenius; Yumeng Wang

Industrial Sector: Biomedical/Healthcare; Data Analysis

Key Words: chronic kidney disease, machine learning, random forests, classification

MSC2020 Codes: 62P10, 92C50

† Corresponding Author: `jamiree@ucsb.edu`

Contents

1 Vironix MPI 2023 Requests

MPI 2023: Fundamental questions

(1) What collection of patient lab data, biometrics readings, symptoms, and baseline health factors indicate acute (rapid) and chronic (slower) deterioration of chronic kidney disease?

(2) Given a set of patient data describing the different temporal stages of CKD deterioration, can we predict when a patient's health state is at high risk for degeneration to a new state (i.e., stage 2 to stage 3 or stage 3 to end-stage renal disease)?

(3) Can we correlate estimated glomerular filtration rate and creatinine diagnostic levels to other observable patient health data (biometrics, symptoms, other co-morbidities, etc.) to maximize the efficacy of remote monitoring when lab data isn't available?

(4) Can we classify (clustering or otherwise) a set of health states that are at the highest risk of acute and chronic health deterioration (predict hospitalization from Hong data set using existing feature data and identify high-risk patients based on stage level from UCI data set)?

(5) Can we build an analytic model to predict patient scenarios indicative of a mild/severe presentation of heart failure?

(6) What are the performance differences between models built by various machine-learning classifiers? How do machine-learned classifiers compare to analytical approaches?

2 Background

Proper kidney function is critical for maintaining overall good health. The kidneys not only remove wastes, toxins, and excess fluid, but they also help control blood pressure, stimulate the production of red blood cells, maintain the health of the bones, and regulate blood chemicals that are essential to life. According to the Center for Disease Control and Prevention (CDC), more than one in seven American adults are estimated to have chronic kidney disease (CKD). CKD is a condition in which the kidneys are damaged and cannot filter blood as well as they should. Because of this, excess fluid and waste from blood remain in the body and may cause other health problems, such as heart disease and stroke. 15% of US adults are estimated to have chronic kidney disease, about 37 million people [2]. CKD also contributes to additional health concerns such as anemia, increased occurrence of infections, low calcium levels, high phosphorous levels, high potassium levels in the blood, depression, and loss of appetite. CKD has varying levels of severity which are irreversible. It usually worsens over time though treatment has been shown to slow progression. If left untreated, CKD can progress to kidney failure and early cardiovascular disease. Kidney failure treated with dialysis or kidney transplant is called End-Stage Renal Disease (ESRD).

3 Introduction

Chronic kidney disease is a condition caused by damaged kidneys filtering blood abnormally that worsens over time. Usually, the kidneys are responsible for filtering waste and fluid out of the blood to cycle to the rest of the body. Proper kidney function is essential to good health as it allows for stable blood pressure, healthy bones, and continual production of red blood cells.

When the kidneys are damaged, waste, fluid, and toxins can build up in your blood/body and cause severe damage to your health. Damage to the kidneys cannot be reversed and can lead to kidney failure, also known as end-stage renal failure. Today, more than one in seven Americans are estimated to have CKD, most unknowingly. Anyone can get CKD, but those with diabetes, high blood pressure, and heart disease are more likely to develop it, making Black Americans, Native Americans, Asian Americans, and Hispanics highly affected. However, if doctors can detect the damage early and stop further kidney degeneration, patients can manage their symptoms to lead relatively normal lives. Thus, there are copious efforts to identify the progression of CKD through its five stages of degeneration.

4 Biological Background

4.1 Pathology of CKD Degeneration

Chronic kidney disease is usually asymptomatic until it progresses to the later stages. It is, therefore, not defined as CKD until there has been kidney damage present for more than 3 months / 90 days. This damage is currently diagnosed through blood and urine test results.

Kidney dysfunction is commonly measured by the glomerular filtration rate (GFR) of less than 60 mL/min/1.73 m^2 in the blood and increased urinary albumin excretion. CKD has been categorized into 6 stages based on a patient's GFR as follows:

$$\begin{array}{lllllll}
\text{Stage 1:} & 90\ \text{mL/min} & \leq & \text{GFR} & & \\
\text{Stage 2:} & 60\ \text{mL/min} & \leq & \text{GFR} & < & 90\ \text{mL/min} \\
\text{Stage 3a:} & 45\ \text{mL/min} & \leq & \text{GFR} & < & 60\ \text{mL/min} \\
\text{Stage 3b:} & 30\ \text{mL/min} & \leq & \text{GFR} & < & 45\ \text{mL/min} \\
\text{Stage 4:} & 15\ \text{mL/min} & \leq & \text{GFR} & < & 30\ \text{mL/min} \\
\text{Stage 5:} & & & \text{GFR} & \leq & 15\ \text{mL/min}
\end{array}$$

From Ilyas [8], we have the following Modification of Diet in Renal Disease (MDRD) formula to estimate Glomerular Filtration Rate (GFR) depending on age, gender, and serum creatine (SCr):

$$\text{GFR} = 175\,\text{SCr}^{-1.154}\text{age}^{-0.203}[0.742\ (\text{if female})], \tag{4.1}$$

where the "if female" bracket represents selecting 1 if the patient is male or 0.742 if the patient is female. Note that for this equation and the remaining GFR equations to be presented, $\kappa = 0.7$ for females and 0.9 for males. Another formula is alluded to:

$$\text{GFR} = 141 \min\{\text{SCr}/k, 1\}^{\alpha} \max\{\text{SCr}, 1\}^{-1.209} 0.993^{\text{age}}[1.018\ (\text{if female})]. \tag{4.2}$$

Since race is correlated to CKD, we can include this feature in the GFR calculation by the following equation:

$$\text{GFR} = 141 \min\{\text{SCr}/k, 1\}^{\alpha} \max\{\text{SCr}, 1\}^{-1.209} 0.993^{\text{age}}[1.018\ (\text{if female})]\cdot[1.159\ (\text{if Black})] \tag{4.3}$$

where the "if Black" bracket represents selecting 1 if the patient is not Black or 1.159 if the patient is Black. From all of these formulas, it is clear that the higher the SCr, the lower the GFR, and the higher the stage of CKD.

While the disease presents asymptomatically initially, as the stages progress, more symptoms become noticeable. These symptoms can be triggered by certain medications, dehydration, infections, unstable BP, and drug abuse. As the kidneys stop filtering fluid and waste properly, patients may experience:

- muscle cramps
- swelling of legs, ankles, or feet
- increased urination
- insomnia
- depression
- lost appetite

- congestive heart failure (CHF) symptoms

These worsening symptoms make early CKD diagnosis and intervention imperative. Unfortunately, GFR is a measurement obtained only through lab tests, making diagnosis and tracking inaccessible to most Americans. Therefore, many researchers have investigated other health features that may aid CKD diagnosis and dysfunction detection.

5 Mathematical Background

Chronic kidney disease is a complex disease that impacts multiple organ systems throughout the body. Diagnosis and detection are often made through lab data collection and analysis. Mathematical modeling can allow researchers to quantitatively represent multiple components and scales of a system and investigate the dynamic behavior of these components and their interactions over time under various conditions. Thus, many researchers have been employing math modeling techniques to aid in CKD progress prediction such as analytical methods, numerical methods, and machine learning algorithms.

5.1 Machine Learning Methods of Analysis

Kidney is known for its very complex structure and the underlying biology mechanisms driving CKD is very limited as described in Section 4. Applying machine learning to complex data has been widely encouraged. In biology, the goals are often predicting accurately the underlying biological processes of interest when experimental data is not complete, or explaining biological processes [5]. A meta-list of studies which have applied machine learning algorithms using a CKD data set is recorded in [4]. There are two broad types of tasks of machine learning: Supervised and unsupervised learning. The former is used with labeled data that means both inputs/features and outputs are available. Unsupervised learning can be used when data is not labeled. We can assess relationships, structure and characteristics from the data using both types of learning. Regression and classification are considered supervised learning, while clustering is unsupervised. For regression, we used eGFR as our output value to be predicted. We also considered the five stages as a label when performing classification. We dropped the stages and tries to recover them by teaching the model to identify patterns when clustering.

5.2 Cluster analysis via quadratic programming

Given that we possess N measurements for every patient, we normalize them for being in the range $[0, 1]$ and treat as a point in the N-dimensional space. Depending on the CKD severity, the patients are divided in 5 stages, which form the clusters in the N-dimensional space.

Hypothesis 1: The distance between each successive stage (cluster) is the same. In other words, successive stages are equidistant concerning special metrics.

Hypothesis 2: The parameters of the metrics identify the significance of the corresponding measurements.

In the N-dimensional space, the five points are formed from the five stages of CKD. Each feature has N measurements (coordinates), each of which has a cluster which is the average of the sizes of each individual for that specific parameter. We form five standards of the known representatives from each stage: X_1, X_2, X_3, X_4, X_5—the center points of the clusters. Let the weighed "distance" between clusters j and k be as follows

$$d_{jk}^2 = d^2(X_j, X_k) = \sum_{i=1}^{N} w_i \cdot (x_{ji} - x_{ki})^2,$$

where x_{ji} is the i-th component of point X_j. Weight parameters w_i are yet unknown. For the usual Euclidean metrics, $w_i = 1$ and their sum is N. Then, on vector w, we impose the following natural constraints

$$\sum_{i=1}^{N} w_i = N, \qquad w_i \geq 0. \tag{5.1}$$

Following Hypothesis 1, we should minimize, with respect to w, the following functional.

$$F(w) = (d_{12}^2 - d_{23}^2)^2 + (d_{23}^2 - d_{34}^2)^2 + (d_{34}^2 - d_{45}^2)^2 = \sum_{j=1}^{3} \left(\sum_{i=1}^{N} w_i \cdot z_{ji} \right)^2 \tag{5.2}$$

where

$$z_{ji} = (x_{ji} - x_{j+2,i}) \cdot (x_{j,i} - 2x_{j+1,i} + x_{j+2,i}).$$

In fact, as we see

$$d_{12}^2 - d_{23}^2 = \sum_{i=1}^{N} w_i \cdot \Big((x_{1i} - x_{2i})^2 - (x_{2i} - x_{3i})^2 \Big) = \sum_{i=1}^{N} w_i (x_{1i} - x_{3i}) \cdot (x_{1i} - 2x_{2i} + x_{3i}),$$

and the substitution yields us a formula for z_{ji} in (5.2).
In order to minimize functional F with constraints

$$w_i \geq 0 \quad \text{and} \quad g(w) = \sum_{i=1}^{N} w_i = N, \tag{5.3}$$

we use Lagrange multipliers, $\nabla F = \lambda \nabla g$, and obtain

$$(\nabla F)_k = 2 \sum_{i=1}^{N} y_{ki} w_i = \Lambda$$

where $y_{ki} = z_{1i} z_{1k} + z_{2i} z_{2k} + z_{3i} z_{3k}$ and Λ is a constant N-dimensional vector $\Lambda = \langle \lambda, \lambda, \ldots, \lambda \rangle$. Symmetric positive entries y_{ki} form $N \times N$ matrix Y, and we arrive at the linear system

$$Yw = \Lambda$$

with constraints as in (5.3). Since λ is undetermined, we can solve the system $Yw = \Lambda$ for, say, $\lambda = 1$ burdened with the only constraint $w \geq 0$. Then, the solution can be normalized to meet the constraint $w_1 + w_2 + \ldots + w_N = N$.

Given our described set up, we note that it is still an open problem whether the solution w has a distinguished sign, i.e., whether all w_i are either positive or negative. Moving on, let $w = \operatorname{argmin} F(w) \in \mathbb{R}^N$ with (5.1) taken into account. Given additional normalized patient data x, we measure $d(x, X_j)$ for $1 \leq j \leq 5$ and assign x to the nearest cluster (stage), say, X_2. Then, we recalculate the average X_2 and, an instant later, weight vector w. Then we repeat this process as many times as we have patient data. The found coordinates of vector w will demonstrate the significance level for the corresponding measurement and, thus, the influence of every measurement on the patient's condition. As such, we can find which parameters are most significant in determining the stage of CKD for an individual and the additional parameters essential for assigning steps.

Remark. We expect that for every triangle $X_iX_jX_k$ with $i < j < k$, the longest side is X_iX_k.

To get a head start on the computational optimization approach, we began with an attempt at reformulating the problem as a linear program. In essence, we tried altering the quadratic objective function $F(w)$ to obtain the linear objective function $F_L(w)$ but we have not done so successfully.

5.3 Feature Selection for Predictive Importance/Correlation

GFR and creatinine levels are the most essential features in CKD stage prediction. This should be somewhat intuitive since the stage of CKD is defined by GFR and since creatinine levels are used in a formula to calculate GFR. Since we know this, we are interested in finding features that do not include GFR and creatinine to see how healthy models perform without them. From Kikuchi et al. [9] we know that low Body Mass Index (BMI) and serum albumin levels are correlated with CKD progression. Albumin is a protein made by the liver, and albumin levels can be measured in both blood and urine. Since BMI can be easily calculated, and albumin can be roughly estimated with some at-home tests, these features may help be included. However, it is essential to note from [1] that our understanding of how High BMI correlates with CKD is inconclusive.

The advancement of biomedical sensors and network technologies provides a widespread progression in the Internet of Things (IoT) field as a system where intelligent medical devices with unique identifiers can be connected for early diagnosis of severe conditions such as CKD. A study by [7] uses a five-phase system including: (1) Collecting necessary data with biomedical sensors and innovative multimedia medical devices, (2) Preprocessing collected data, (3) Feature selection, (4) CKD prediction process based on the existing classification methods, (5) Performance analysis via sensitivity, accuracy, and specificity factors. The features used in their prediction model included twenty-six vital signs that influence CKD, plus five clinical features (frequent urination, foot swelling, insomnia, chest pain, and weight loss) as additional nominal variables. The extracted rules for kidney functionality from their classification results included the following features (other than GFR and serum creatinine) as necessary: age, blood pressure, weight loss, foot swelling, insomnia, chest pain, frequent urination, diabetes mellitus, red blood cells level, albumin, white blood cells count, anemia, presence of bacteria, potassium, and specific gravity. The resulting decision tree classifier produces accuracy results of 96–97% for three data sets.

6 Data Processing

6.1 Data Cleaning

There are common paths for imputation, each with its own set of implications. Here we discuss these implications and explain our chosen imputation methods.

- **Dropping Unknown Values:** If we were to drop all rows with unknown values, we would not only end up throwing away a lot of valuable data but would also (likely) end up with a significantly smaller dataset or throw away all of our data altogether. Even after dropping columns that may be less significant, we might still end up tossing about 70% of the dataset [12].
- **Imputation/Interpolation/Linear Regression:** We are cautious with using imputation with medical data, as we do not wish to make any assumptions about personal data that may be unfounded [12].
- **Quantization on Thresholds:** The goal of quantizing a feature in one of our datasets is to approximate the feature in question by restricting the amplitude of the values to a prescribed set of values [12].

7 Exploratory Data Analysis

7.1 Correlation Analysis

In dataset [8], in addition to the stage of CKD for each patient, other parameters are also provided. However, using all of these parameters to help predict a patient's stage of CKD is quite cumbersome. Instead, we only want to use a small subset of these parameters. In order to determine which parameters would be most useful for the prediction, we want to determine if there is any significant correlation between any of the parameters and the CKD stage. We do this through two methods: scatter plots and a correlation matrix.

7.1.1 *Scatter Plots*

For the numerical parameters in [8], we form scatter plots of the patients' assigned value to that parameter against their designated stage of CKD. As a note, for any missing values of a parameter, i.e., labeled as a "?", we convert this to NaN. Further, for each parameter, we remove any NaN values. Note that for scatter plots for sodium and potassium, we have removed a few outliers in order to more accurately capture the trend of the data in comparison to CKD stage.

From Figure 1, we see possible positive correlations between blood glucose random (bgr), blood urea (bu), potassium (pot) and CKD stage. Similarly, we see negative correlations between hemoglobin (hemo), red blood cell count (rbcc), packed cell volume (pcv), sodium (sod), and CKD stage. On the other hand, we are seeing no noticeable trend between age and CKD stage. Also, for blood pressure (bp), it is unclear if there is a weak positive correlation between it and CKD stage. Further, it is unlikely that we can use age or blood pressure (bp) to help predict CKD stage. On the other hand, our results suggest that decreasing levels of hemo, rbcc, pcv, and/or sod are possible indicators that a patient is in the later stages of CKD. Increasing levels of bgr, bu, and/or pot may

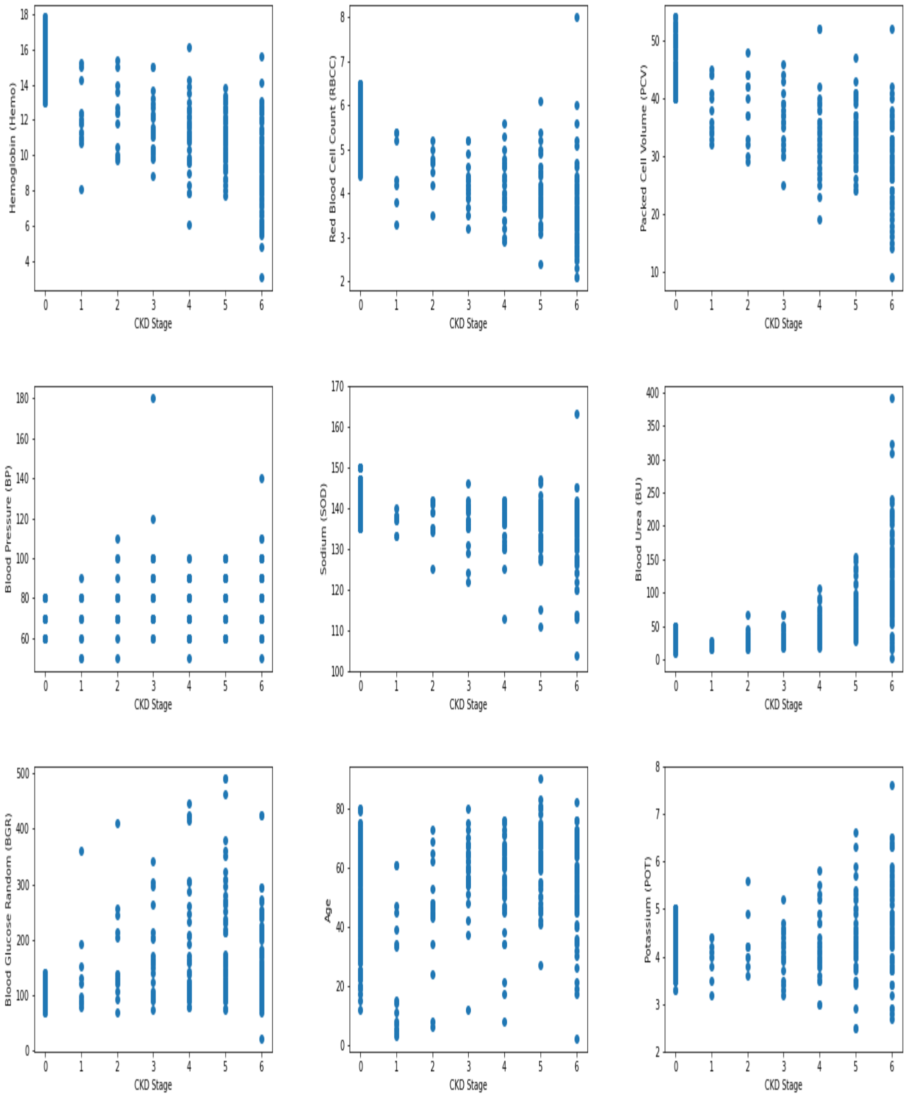


Figure 1. Scatter Plots to Determine Correlation Between Variables and CKD Stage in [8].

also be indicators of this. Thus, hemo, rbcc, pcv, sod, bgr, bu, and/or pot may be good candidates for predicting the CKD stage of a patient.

7.1.2 *Correlation Matrix*

In the previous subsection, we found that some variables are correlated with the CKD stage. However, these scatter plots can only be used for the numerical data and we want to be able to explore the nominal data as well. Thus, we form a correlation matrix for all the

	Raw value	Cleaned value
rbc/pc	normal	1
	abnormal	0
pcc	present	1
	not present	0
race	African American	1
	not African American	0
sex	m	1
	f	0
htn/dm/cad/pe/ane	yes	1
	no	0
appet	good	1
	poor	0
class	not ckd	stage 0
	stage 1	stage 1
	stage 2	stage 2
	stage 3A	stage 3
	stage 3B	stage 4
	stage 4	stage 5
	stage 5	stage 6

Table 1: Data cleaning of [8] for correlation analysis.

parameters. Correlation matrices determine the correlation coefficients between any two variables which can be indicative of possible relationships between the variables. These coefficients take on values between -1 and 1. Positive coefficients indicate that an increase in one variable can lead to an increase in the other. Negative coefficients indicate that an increase in one variable can lead to a decrease in the other. The closer the magnitude of the coefficient is to 1, the stronger the relationship. Forming a correlation matrix can help us determine which variables would be good candidates for predicting the CKD stage. In addition, the correlation matrix can help us determine which parameters are highly correlated with each other. This phenomenon is known as multicollinearity. For example, if two parameters are highly correlated with CKD stage but are also highly correlated with each other, it can make it more difficult to determine the individual effects of using these parameters in a model to predict the CKD stage. Thus, we would only need to select one of these parameters in our model. We make use of Python's corr. function in order to produce the correlation matrix which uses Pearson's coefficient and ignores any NaN values. As a note, we consider any two variables to be highly correlated if their correlation coefficient has a magnitude greater than or equal to 0.5 [6, 11].

From Table 2, we see that hemoglobin (hemo), red blood cell count (rbcc), and packed cell volume (pcv) are the variables most highly negatively correlated with CKD stage. In particular, the correlation coefficients have magnitudes in the range 0.71–0.79. At the

Positive Correlation	Negative Correlation
Albumin (al): 0.65	Hemoglobin (hemo) : -0.79
Hypertension (htn): 0.63	Packed cell volume (pcv): -0.78
Blood urea (bu): 0.6	Red blood cell count (rbcc): -0.71
Diabetes (dm): 0.56	Specific gravity (sg): -0.66
	Red blood cell (rbc): -0.56
	Pus cell (pc): -0.52

Table 2: Significant correlations of other variables to CKD stage (Figure 2); red/blue = variables highly correlated with each other.

same time, these three variables are highly positively correlated with each other with correlation coefficients of magnitudes 0.78 or 0.9. Given that hemoglobin is carried in the blood through red blood cells and packed cell volume measures the percentage of blood that is occupied by red blood cells, it is not surprising that these variables are highly positively correlated with each other. Further, given that hemoglobin carries vital oxygen to organs such as the kidneys, we can see why hemoglobin (and rbcc and pcv) is highly negatively correlated with CKD stage.

In addition, diabetes (dm), hypertension (htn), blood urea (bu), albumin (al), specific gravity (sg), red blood cell (rbc), and pus cell (pc) are moderately correlated with CKD stage where the former three variables are positively correlated with CKD stage while the latter three are negatively correlated with CKD stage. In particular, their correlation coefficients have magnitude in the range 0.52–0.66. At the same time, diabetes and hypertension are positively correlated with each other with correlation coefficient 0.61. Given that illnesses can cause other problems in the body, it can be seen why diabetes and hypertension are positively correlated with CKD stage and with each other. Similarly, given that variables like blood urea and albumin provide blood measurements indicative of kidney function, we can intuitively see the correlations of these variables of CKD stage.

Overall, our results appear to support our conclusions from the previous subsection as we see negative correlations between hemoglobin, rbcc, and pcv and CKD stage as we do in Figure 1. Similarly, we see positive correlations between blood urea and CKD stage. However, unlike in the previous subsection, there is not a high correlation between sodium or potassium and CKD stage. Thus, in the future, if we were to use a machine learning algorithm to predict the CKD stage of a patient, we will only consider variables in Table 2. In particular, we would focus on using one of the three variables, hemoglobin, rbcc, pcv, for such an algorithm due to their high correlation with CKD stage.

7.1.3 *K-means*

K-means methods can be used to find how subsets of a gathered dataset can be strongly correlated with each other in terms of a subset of desired variables. We used K-means algorithms on the Hong dataset using the parameters we found correlated most strongly with eGFR (excluding creatinine) and found that K-means clusters did not seem to

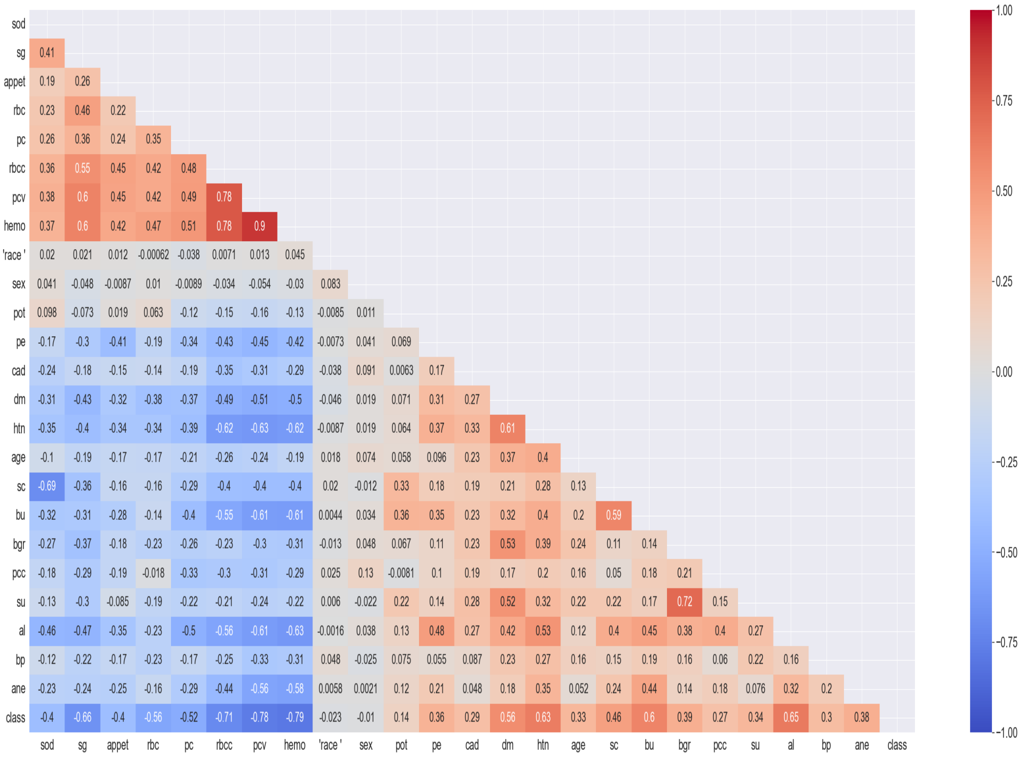


Figure 2. Correlation matrix of data from [8]

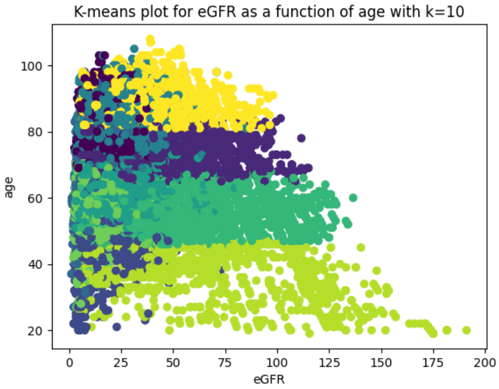


Figure 3. K-means with significant parameters and $k = 10$.

correlate to classes of CKD (or equivalently eGFR). Here is one example of one such plot.

7.2 Random Forest

Random forest algorithms (also known as random decision forests) are ensemble learning methods for classification, regression, and other tasks that operate by constructing a

Feature	Weight
Blood Urea Nitrogen Level*	0.253329
Anion Gap*	0.201570
Chloride Level*	0.180366
Red Blood Count*	0.092631
Hemocrit Level*	0.070413
Chloride Level*	0.045939
Bicarbonate Level*	0.026342
Potassium Level*	0.024307
Sodium Level*	0.022485
Platelets Level*	0.020696

* median measurement used.

Table 3: Key features (including lab measurements) used to predict CKD stages, and their relative weights.

Feature	Weight
Age	0.195009
Number of In-patient Admissions**	0.096965
Number of Emergency Department Visits**	0.092732
Systolic Blood Pressure*	0.089689
Race: White or Caucasian	0.075582
Albumin*	0.057300
Anemia	0.056261
Race: Black or African American	0.048469
Pulse* 0.047619	

* median measurement used.

** within the past year.

Table 4: Key features (including lab measurements) used to predict CKD stages, and their relative weights.

multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over-fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees, which Vironix has used in the past for CKD prediction. However, performance can be affected by data characteristics (Wiki).

7.3 Gradient Boosting

Gradient boosting is a supervised machine-learning technique similar to a random forest. Like other boosting algorithms, the model is built stage-wise with decision trees.

Feature	Weight
Age	0.05650321
Number of In-patient Admissions**	0.040257808
Race	0.048045773
Systolic Blood Pressure*	0.009659265
Anion Gap *	0.36311668
Blood Urea Nitrogen Level*	0.120267235
Calcium Level*	0.0552984
Chloride Level*	0.21430616
Hematocrit Level*	0.01613212
Potassium Level*	0.07641336

* median measurement used.

** within the past year.

Table 5: Key features (including lab measurements) used to predict CKD stages, and their relative weights.

Feature	Weight
Age	0.10409631
Albumin Level*	0.04753141
Anemia	0.20151092
Diastolic Blood Pressure Level*	0.06589365
Glucose Level*	0.03212508
Number of In-patient Admissions**	0.05681709
Number of Emergency Department Visits**	0.2665141
Race	0.1642751
Systolic Blood Pressure*	0.061236244

* median measurement used.

** within the past year.

Table 6: Key at-home features (excluding lab measurements) used to predict CKD stages and their relative weights.

However, it allows for a loss function to be minimized. This generalization means that it performs very well and tends to outperform random forests. We use XGBoost, which is a library that implements boosted trees in an optimized way. Furthermore, XGBoost automatically handles missing values in the dataset: branch directions are learned during training, which is another advantage of using gradient boosting over random forest algorithms. Using XGBoost, we find that the most significant features to predict CKD stages in general (including lab results) in Table 5, and the most significant features excluding lab measurements (just at-home measurements) can be seen in Table 6.

Feature	Description	Type	Lab Generated
age	Age of patient	Numeric	0
acrenlfail	History of Acute and unspecified renal failure	Binary	0
albumin	Median Albumin levels	Numeric	0
anemia	History of Deficiency and other anemia	Binary	0
cardiaarrst	History of Cardiac arrest and ventricular fibrillation	Binary	0
cc breathingdifficulty	Chief complaint breathing difficulty	Binary	0
cc breathingproblem	Chief complaint breathing problem	Binary	0
cc chestpain	Chief complaint chest pain	Binary	0
cc chesttightness	Chief complaint chest tightness	Binary	0
cc edema	Chief complaint edema	Binary	0
cc fever	Chief complaint fever	Binary	0
cc footswelling	Chief complaint foot swelling	Binary	0
coronathero	History of Coronary atherosclerosis	Binary	0
dbp	Diastolic Blood Pressure	Numeric	0
diabmelnoc	History of Diabetes mellitus without complication	Binary	0
diabmelwcm	History of Diabetes mellitus with complications	Binary	0
ethnicity	Ethnicity	Categorical	0
gender	Gender	Binary	0
glucose	Median Glucose levels	Numeric	0
htncomplicn	History of Hypertension with complications	Binary	0
kidnyrnlca	Cancer of kidney and renal pelvis	Binary	0
n admissions	number of in-patient admissions in the past year	Numeric	0
n edvisits	number of ED visits in the past year	Numeric	0
nauseavomit	Nausea and vomiting	Binary	0
o2 device	At home oxygen concentrator	Binary	0
otdxbladdr	History of Other diseases of bladder and urethra	Binary	0
otdxkidney	History of Other diseases of kidney and ureters	Binary	0
othheartdx	History of Other ill-defined heart disease	Binary	0
pulse	Pulse rate	Numeric	0
race	Race	Categorical	0
resp	Respiration rate	Numeric	0
sbp	Systolic blood pressure	Numeric	0
temp	Temperature (F)	Numeric	0
triage vital hr	Heart rate	Numeric	0
whtblooddx	History of Diseases of white blood cells	Binary	0

Table 7: At-home features identified as important in detecting CKD stages.

7.4 Visualizations of Health Parameters as a Function of the Disease Stage

The following variables were identified as important features in Dataset 2 using the insights from literature study, correlation analysis, and initial machine learning models. The features were then grouped into two sets: At-home and Lab Generated. The At-home variable set is identified based on the assumption that if a patient has an additional condition, they would know that and hence would be able to identify themselves as having that concern based on their medical history. Tables 7 and 8 show the features and whether or not they were lab generated using a binary indicator, where Lab Generated=1 would indicate that the feature is in the Lab Generated set and 0 in the At-home variable set.

In trying to study the characteristics of each stage of CKD, we created some visualizations to show the severity of CKD exhibits in the form of some at-home features.

Feature	Description	Type	Lab Generated
aniongap	Anion gap of blood	Numeric	1
bun	Blood urea nitrogen	Numeric	1
calcium	Calcium levels	Numeric	1
chloride	Chloride elvels	Numeric	1
c02	Carbon dioxide in blood	Numeric	1
creatinine	Creatinine	Numeric	1
globulin	Globulin level in blood	Numeric	1
hematocrit	Hematocrit	Numeric	1
hemoglobin	Hemoglobin	Numeric	1
magnesium	Magnesium Levels	Numeric	1
mch	Mean Corpuscular Hemoglobin	Numeric	1
mcv	Mean Corpuscular Volume	Numeric	1
mpv	Mean Platelet Volume	Numeric	1
platelets	Platelets count	Numeric	1
potassium	Potassium Levels	Numeric	1
rbc	Red blooc count	Numeric	1
sodium	Sodium levels	Numeric	1
wbc	White blood count	Numeric	1

Table 8: Lab generated features identified as important in detecting CKD stages.

Figure 4 shows that as age progresses, eGFR levels tend to go down for both males and females, however the decline for males is slightly stronger. Figure 5 shows that heart rate of patients tends to vary for lower levels of eGFR and systolic blood pressure also increases as eGFR goes lower. Temperature readings also show higher variation at lower eGFR levels for both males and females as depicted in Figure 6.

The difference across the stages of CKD can be seen in reference to the prevalence of anemia, number of emergency room visits, and number of in-patient admissions in the last year from Dataset 2 in Figures 7, 8, and 9 respectively. An interesting observation across these charts is that males generally show more symptoms of severity of CKD but females show higher prevalence of anemia as CKD deteriorates. Figure 9 also illustrates how the transition from Stage 2 to Stage 3A shows a spike in hospital admissions for the same patient, highlighting the need for intervention during Stage 2 to prevent escalation of the disease. Note that the violin plot used here normalizes values.

8 CKD Predictions

8.1 Identification of patient states that indicate a high risk of transition from one state to another

We performed binary classifications for each consecutive pair of stages ($stage_i$ and $stage_{i+1}$). There are 5 pairs in total: stage 1 to stage 2, stage 2 to stage 3A, stage 3A to stage 3B, stage 3B to stage 4, and stage 4 to stage 5. The patient states that indicate a high risk of transition from one stage to another are given by the most im-

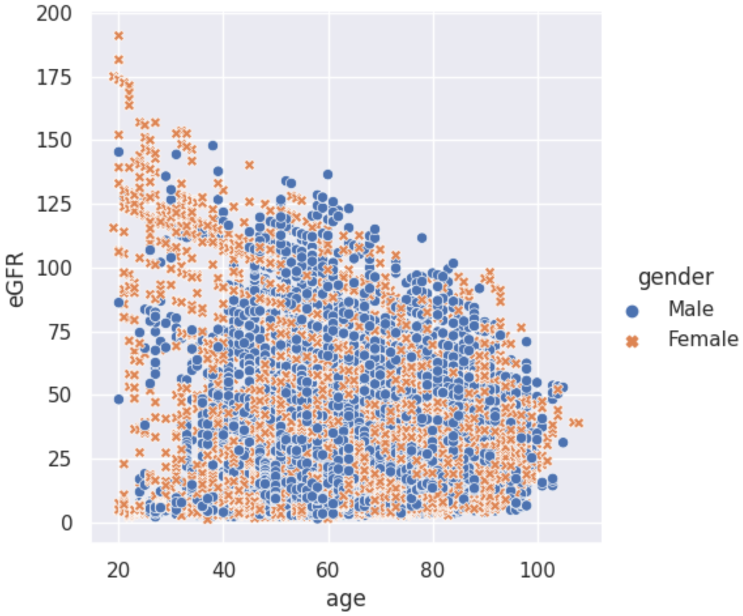


Figure 4. eGFR Levels in Dataset 2 against Age and Gender of patients.

portant features when classifying the current stage to the next. LASSO (least absolute shrinkage and selection operator) is a penalized modeling approach and has been used in medical settings [10]. It finds the set of coefficient estimates that best fit the data where up to a certain number of nonzero coefficients are only allowed. Support vector machine is a supervised learning method used for classification and regression models. For classification, it is call support vector classification which finds a separation plane where the distance between the data points from each class to the plane is maximized [3]. We carried out LASSO on linear support vector classification to classify each consecutive pair of stages in dataset 1 [8]. After cleaning the data as elaborated in Section 6.1, two consecutive stages are considered as the response variable to be separated including all the features. For each consecutive stages pair, we performed a validation set approach for model performance that is we randomly divided the dataset into half for training set and validation set. We also record the accuracy assessment with 100 repetitions to get 100 misclassification error rates and averaged them for assessing the prediction performance as the dataset is small.

Table 9 summarizes the accuracy of the model on each pairwise consecutive stages. Stage 4 to stage 5 has the highest accuracy 0.7 compared to other pairs (0.6 accuracy). That means, classifying patients from stage 4 to stage 5 is easier than identifying patients from among one lower stage to another.

Table 10 records the five most important features for each consecutive pairwise stages and Figure 10 displays the importance rates of all the features. A high risk of stage 1 patients transitioning to stage 2 depends on the appetite, race, pus cell clumps and potassium level. Race appears as important when identifying high risks of transitioning from one state to another for all five pairs. Red blood cells count, race, and pus cell

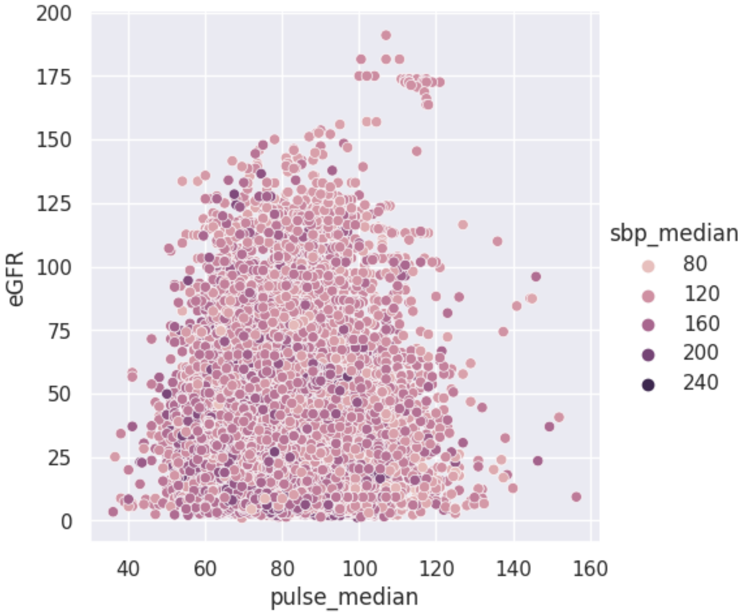


Figure 5. eGFR Levels in Dataset 2 against Pulse and Systolic blood pressure of patients.

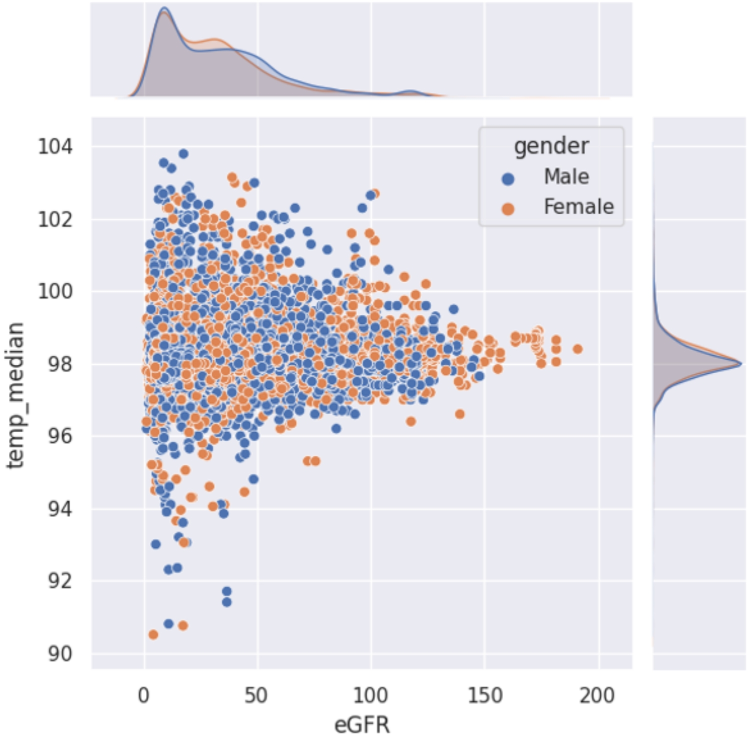


Figure 6. eGFR Levels in Dataset 2 against Temperature.

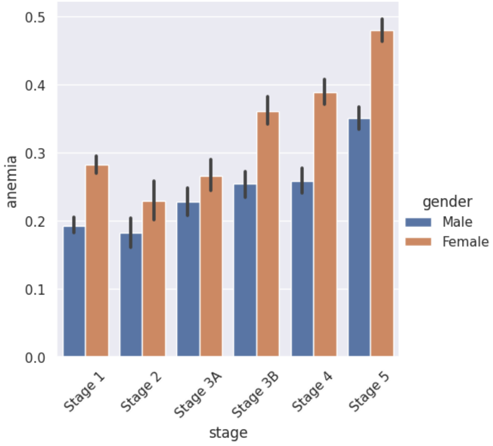


Figure 7. Prevalence of Anemia at each CKD Stage in Dataset 2.

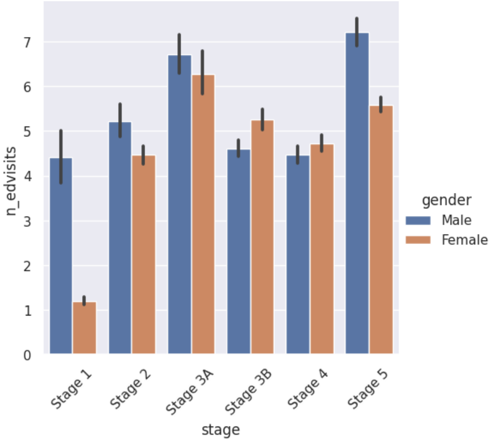


Figure 8. Number of ER visits at each CKD Stage in Dataset 2.

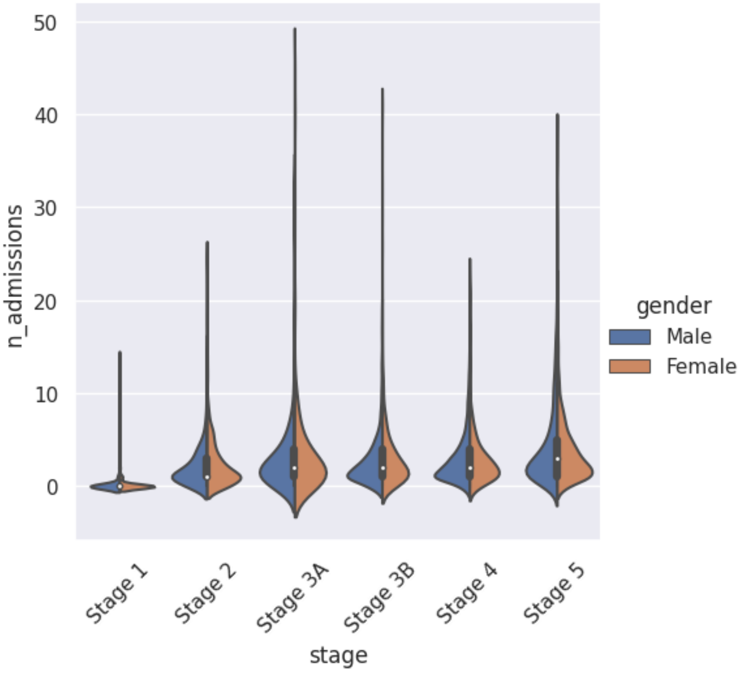


Figure 9. Number of inpatient admissions at each CKD Stage in Dataset 2.

	1 to 2	2 to 3A	3A to 3B	3B to 4	4 to 5
LinearSVC	0.6	0.6	0.6	0.6	0.7

Table 9: Accuracy level of linear support vector classifier for each consecutive pair of stages.

Stage 1 to 2	Stage 2 to 3A	Stage 3A to 3B	Stage 3B to 4	Stage 4 to 5
appetite	race	race	race	race
coronary artery disease	anemia	pus cell clumps	red blood cell count	blood glucose random
race	pedal edema	red blood cell count	pus cell clumps	pus cell
pus cell clumps	red blood cell count	potassium	diabetes mellitus	coronary artery disease
potassium	coronary artery disease	anemia	pedal edema	diabetes mellitus

Table 10: Feature importance of all five consecutive pairs of stages with linear support vector classifier.

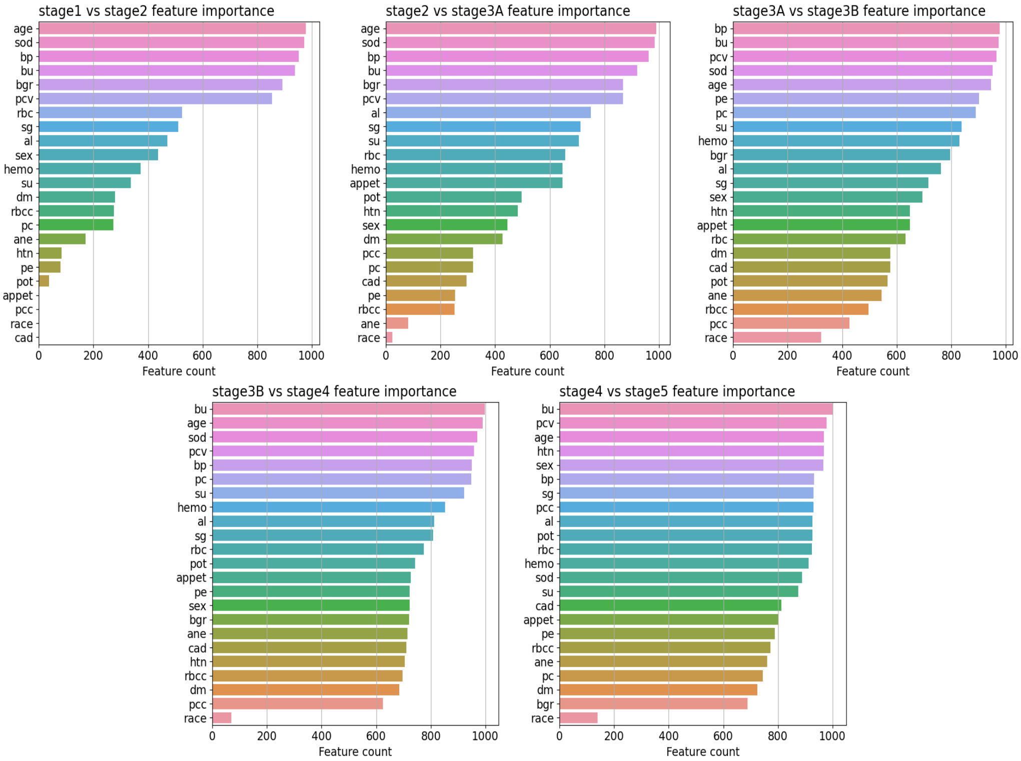


Figure 10. Feature importance for each consecutive pair of stages of dataset 1 [8] based on linear support vector classifier.

clumps are important when assessing risks of transitioning from low stage to high stage of CKD (stage 2 to stage 3a, stage 3a to stage 3b, and stage 3b to stage 4).

For each pair, the accuracy metric and confusion matrix were also recorded for other classification learning, such as, random forest (RF) 7.2, gradient boosting classifier (GB) 7.3, and multi-layer perceptron classifier (MLP) [13]. For each pair we used confusion matrix and predictions performance metrics to select their 'best' model, that is, with high accuracy metric and reasonable confusion matrix (current stage $stage_i$ is misclassified more than the next stage $stage_{i+1}$). Table 11 summarizes the selected model for each pair along with their respective accuracy. Separating each pair using their respective selected model, Figure 11 shows that blood urea and age are important to assess risks of

	Stage 1 to 2	Stage 2 to 3A	Stage 3A to 3B	Stage 3B to 4	Stage 4 to 5
Selected Model	RF	GB	RF	GB	RF
Accuracy	0.6	0.7	0.7	0.7	0.8

Table 11: Selected models of each pair from linear support vector classifier, random forest, gradient boosting classifier, multi-layer perceptron classifier, along with their respective accuracy.

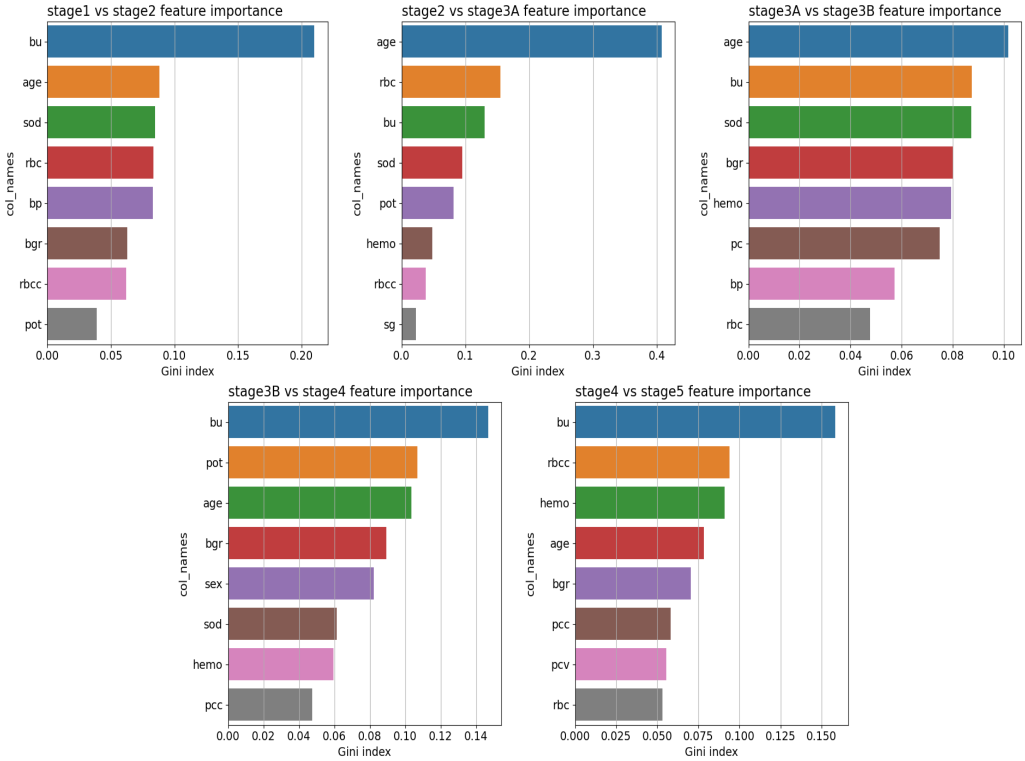


Figure 11. Feature importance for each pair of consecutive stages in [8] based on their selected classification model.

transitioning from one stage to another for all consecutive pairs of stages. Sodium and red blood cells can be useful in distinguishing consecutive pairs of stage 1 to stage 2, stage 2 to stage 3a, and stage 3a to stage 3b. Potassium, hemoglobin, red blood cells count, and blood glucose random are important when assessing risks of transitioning from stage 3b to stage 4 and from stage 4 to stage 5.

In summary, blood urea, age, and race are very important when assessing risks of transitioning from one state of CKD to another, which can all be checked from home. In addition, for low stages (stage 1 and 2), checking for red blood cells and sodium level can be useful when identifying their next stage transitioning risks. For high stages (stage 3a and above) blood glucose random and hemoglobin should be tracked down. The higher the state level is, the many the features that need to be included when assessing the risks of transitioning from one state to another. Fortunately, these features can be mainly tested from home.

8.2 Creatinine prediction

As described in Section 4, a patient's CKD stage is determined by their GFR level. This GFR level is calculated using (4.3). To predict a patient's CKD stage, we first predict a patient's creatinine level and then use the (4.3) and GFR bounds to classify a patient's stage. We use linear regression and random forest regression to predict creatinine levels using the features given in [8].

From the Ilyas data [8], we select the 379/400 patients that have a creatinine reading, use a one-hot encoder for the nominal variables (including NaN values as a separate value to represent missing data), impute the numerical variables with the column mean, and standardise each numerical variable to mean 0 and standard deviation 1.

We run linear regression and random forest regression (with grid search) with the log of creatinine as our dependent variable. Using the log of creatinine eliminates predictions of negative creatinine as well as providing better prediction for low levels of creatinine. In Table 12 we report the mean square error and R2 score for our test and train sets, generated with a 25-75 split. In Figures 12 and 13 we show the true against predicted log creatinine.

We continue with random forest regression and investigate feature importance. A key question from Vironix is whether we can predict CKD degeneration using at-home variables. Our feature importance analysis in Figure 14 reveals blood urea is the most important feature. A naive Amazon search reveals home test kits for blood urea which reveals some promise for Vironix to do at-home prediction of CKD degeneration.

Finally, we use (4.3) to map predicted creatinine to CKD stage. We show the confusion matrix in Figure 15 with an overall accuracy of 66.5% using random forest regression. We see that we perform worst on Stage 1 and 3A and which suggests these are difficult stages to differentiate from others. We over-predict more than under-predict and this is what we would want a model to do, to emulate cautious doctors.

With this small data set, we are able to get good prediction of a patient's creatinine level and reveal an at-home measurement (blood urea) as the most important feature for this prediction. This preliminary regression analysis should give Vironix confidence that they are able to develop models to accurately answer their fundamental questions with CKD and use machine learning to facilitate at-home diagnoses.

For future work, to predict degeneration of CKD, we could use the continuous variable 'GFR' to quantify if a patient is at risk of transitioning into the next stage. Ideally this classification would incorporate temporal data to follow a patient's progression through the disease. However with the current data, one could map the GFR to a continuous 'stage' variable which could be a linear interpolation of Stage 1-5 at the boundary GFR level. Further work could also investigate using neural networks for creatinine prediction as well as restricting the features to only at-home measurable features.

Metric	Linear regression		Random forest	
	Test	Train	Test	Train
MSE	0.202	0.153	0.160	0.213
R2 score	0.733	0.797	0.739	0.772

Table 12: Linear regression and random forest regression mean square error and R2 score for the test (75%) and train (25%) data-sets.

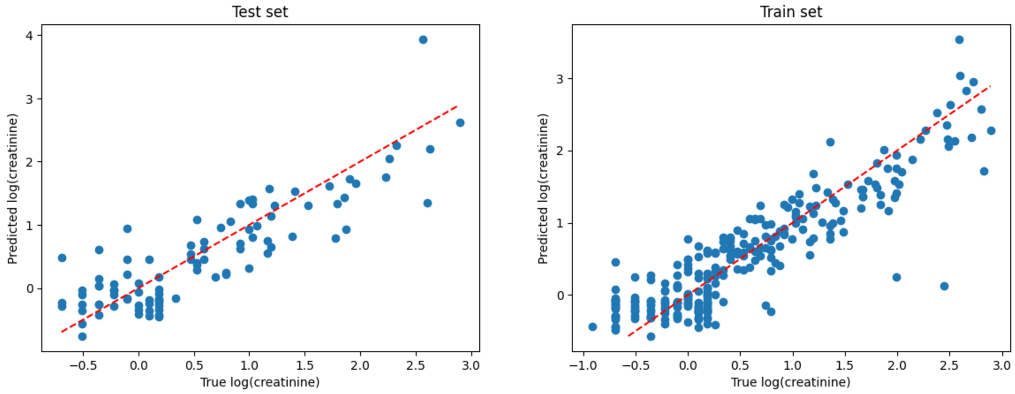


Figure 12. Linear regression results showing the predicted creatinine levels against true creatinine for the test data (left) and train data (right) using a 75-25 train-test split on the Ilyas data [8].

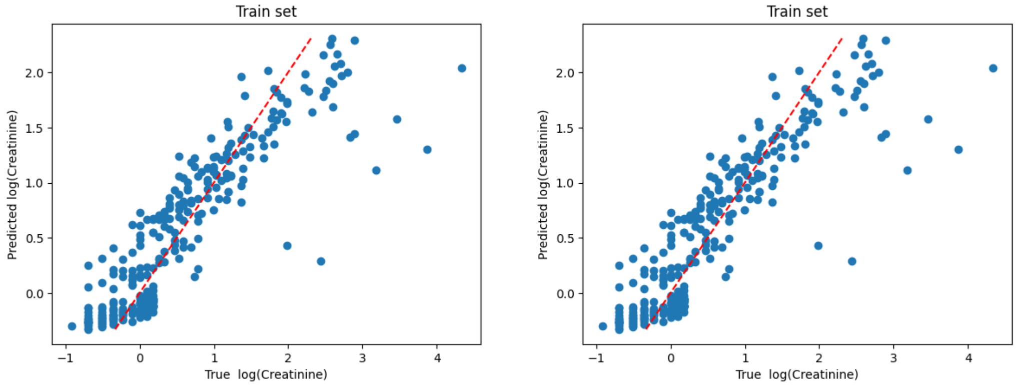


Figure 13. Random forest regression results showing the predicted creatinine levels against true creatinine for the test data (left) and train data (right) using a 75-25 train-test split on the Ilyas data [8].

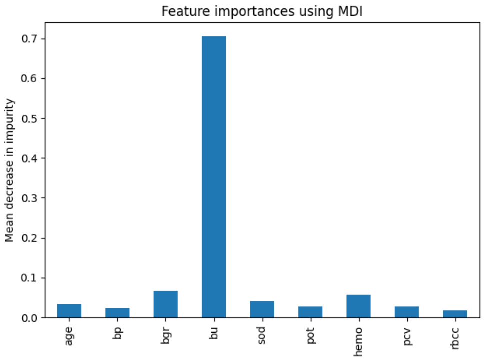
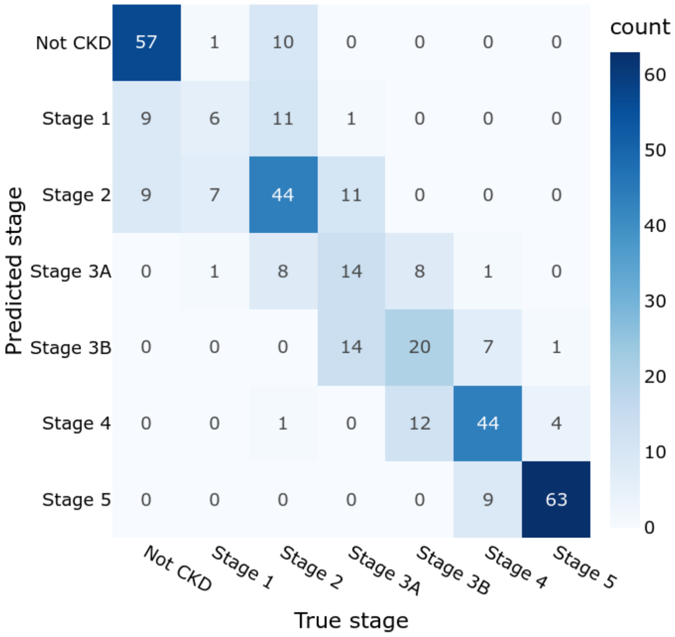


Figure 14. Random forest regression feature importance for numerical variables.



	Not CKD	Stage 1	Stage 2	Stage 3A	Stage 3B	Stage 4	Stage 5
Not CKD	57	1	10	0	0	0	0
Stage 1	9	6	11	1	0	0	0
Stage 2	9	7	44	11	0	0	0
Stage 3A	0	1	8	14	8	1	0
Stage 3B	0	0	0	14	20	7	1
Stage 4	0	0	1	0	12	44	4
Stage 5	0	0	0	0	0	9	63

Figure 15. Random forest regression CKD stage confusion matrix.

8.3 Health states indicative of end-stage renal disease

In Table 13, we compared three different algorithms: random forest, gradient boosting, and a tree-based bagging method. We evaluated their performance using four metrics for all the different stages (weighted): accuracy, precision, recall, and F1 score. We also measured the performance of different algorithms at each stage. Table 13 is an updated table including the evaluation results for each algorithm at various stages for their accuracy, sensitivity (recall, NPV), specificity and precision (PPV).

Algorithm	**Stage**	**Accuracy**	**Sensitivity**	**Precision**	**Specificity**
Random forest	stage 1	0.96	0.77	0.525	0.97
	stage 2	0.90	0.52	0.46	0.94
	stage 3A	0.87	0.70	0.11	0.88
	stage 3B	0.75	0.41	0.64	0.90
	stage 4	0.82	0.47	0.25	0.85
	stage 5	0.82	0.67	0.92	0.95
Gradient boosting	stage 1	0.49	0.30	0.64	0.99
	stage 2	0.47	0.0	–	1.0
	stage 3A	0.46	0.03	0.76	0.99
	stage 3B	0.44	0.089	0.29	0.94
	stage 4	0.45	0.41	0.31	0.77
	stage 5	0.52	0.80	0.39	0.42
Tree-based bagging	stage 1	0.98	0.86	0.83	0.99
	stage 2	0.95	0.72	0.73	0.97
	stage 3A	0.91	0.72	0.60	0.94
	stage 3B	0.87	0.66	0.73	0.93
	stage 4	0.89	0.76	0.73	0.93
	stage 5	0.93	0.88	0.90	0.96

Table 13: Algorithm comparison among stages.

9 Conclusions

We have managed to make significant headway in answering many of the fundamental questions that we set out to answer. Collections of patient lab data, biometric readings, symptoms, and baseline health factors were found to indicate deterioration of CKD. We correlated several metrics to eGFR and creatinine levels and identified several promising metrics to be used for remote monitoring when lab data isn't available. A set of health states were classified as being higher risk of CKD. Performance differences between various machine learning classifiers were provided, but we were unable to provide an analytical approach for which we would compare our classifiers to.

Author Contributions

-MA: ideation, data interpretation.
-AKB: worked with BM on regression etc., write up of Ilyas DS1 stage classification (Section 8.2).
-GSB: performed regression and classification using Ilyas data, figure generation and write up of Ilyas DS1 stage classification with AKB and BM (Section 8.2).
-CC: exploratory data analysis, contributed to section 8.3 (gradient boosting) and section 9.2 (health states indicative of end-stage renal disease).
-JH: conducted literature review, exploratory data analysis, cluster analysis, write-up.
-BM: performed regression and classification experiments with AKB and GSB, figure generation, write up of Section 8.2, proofreading.
-KM: contributed to Section 7.1 (correlation analysis).
-QJ: background, literature review.
-PW: clustering.
-NV: contributed to section 8.
-SAT: Literature review, contributed and write up (Section 8.1)
-YK: background, literature review.

References

[1] D. Barbieri, M. Goicoechea, M. D. Sanchez-Nino, A. Ortiz, E. Verde, U. Verdalles, A. Pérez de José, A. Delgado, E. Hurtado, L. Sanchez-Camara, et al. Obesity and chronic kidney disease progression—the role of a new adipocytokine: C1q/tumour necrosis factor-related protein-1. *Clinical kidney journal*, 12(3):420–426, 2019.

[2] CDC. Chronic kidney disease basics, Feb 2022.

[3] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[4] D. A. Debal and T. M. Sitote. Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, 9(109), 2022.

[5] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, 2022.

[6] A. Hayes. Multicollinearity: Meaning, examples, and faqs, 2023. URL https://www.investopedia.com/terms/m/multicollinearity.asp. Last accessed 15 June 2023.

[7] M. Hosseinzadeh, J. Koohpayehzadeh, A. O. Bali, P. Asghari, A. Souri, A. Mazaherinezhad, M. Bohlouli, and R. Rawassizadeh. A diagnostic prediction model for chronic kidney disease in internet of things platform. *Multimedia Tools and Applications*, 80:16933–16950, 2021.

[8] H. Ilyas, S. Ali, M. Ponum, O. Hasan, M. T. Mahmood, M. Iftikhar, and M. H. Malik. Chronic kidney disease diagnosis using decision tree algorithms. *BMC nephrology*, 22(1):1–11, 2021.

[9] H. Kikuchi, E. Kanda, S. Mandai, M. Akazawa, S. Iimori, K. Oi, S. Naito, Y. Noda, T. Toda, T. Tamura, et al. Combination of low body mass index and serum albumin

level is associated with chronic kidney disease progression: the chronic kidney disease-research of outcomes in treatment and epidemiology (ckd-route) study. *Clinical and experimental nephrology*, 21:55–62, 2017.

[10] S. M. Kim, Y. Kim, K. Jeong, H. Jeong, and J. Kim. Logistic lasso regression for the diagnosis of breast cancer using clinical demographic data and the bi-rads lexicon for ultrasonography. *Ultrasonography*, 37(1):36, 2018.

[11] A. Kumar. Correlation concepts, matrix & heatmap using seaborn, 2022. URL `https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/`. Last accessed 15 June 2023.

[12] M. Marino, J. Lucas, E. Latour, and J. D. Heintzman. Missing data in primary care research: importance, implications and approaches. *Family Practice*, 38(2):199–202, 2021.

[13] T. Windeatt. Ensemble mlp classifier design. In *Computational Intelligence Paradigms: Innovative Applications*, pages 133–147. Springer, 2008.