electronics

MDPI

Article

Diagnosing and Characterizing Chronic Kidney Disease with Machine Learning: The Value of Clinical Patient Characteristics as Evidenced from an Open Dataset

Juan Figueroa [1], Patrick Etim [1], Adithyan Karanathu Shibu [1], Derek Berger [1] and Jacob Levman [1,2,*]

[1] Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada
[2] Nova Scotia Health Authority, Halifax, NS B3H 1V8, Canada
* Correspondence: jlevman@stfx.ca

Abstract: Applying artificial intelligence (AI) and machine learning for chronic kidney disease (CKD) diagnostics and characterization has the potential to improve the standard of patient care through accurate and early detection, as well as providing a more detailed understanding of the condition. This study employed reproducible validation of AI technology with public domain software applied to CKD diagnostics on a publicly available CKD dataset acquired from 400 patients. The approach presented includes patient-specific symptomatic variables and demonstrates performance improvements associated with this approach. Our best-performing AI models, which include patient symptom variables, achieve predictive accuracies ranging from 99.4 to 100% across both hold-out and 5-fold validation with the light gradient boosting machine. We demonstrate that the exclusion of patient symptom variables reduces model performance in line with the literature on the same dataset. We also provide an unsupervised learning cluster analysis to help interpret variability among, and characterize the population of, patients with CKD.

Keywords: chronic kidney disease; machine learning; diagnosis; characterization

check for updates

Citation: Figueroa, J.; Etim, P.; Shibu, A.K.; Berger, D.; Levman, J. Diagnosing and Characterizing Chronic Kidney Disease with Machine Learning: The Value of Clinical Patient Characteristics as Evidenced from an Open Dataset. *Electronics* **2024**, *13*, 4326. https://doi.org/10.3390/electronics13214326

Academic Editors: William M. Mongan and Yu Zhang

Received: 6 September 2024
Revised: 18 October 2024
Accepted: 30 October 2024
Published: 4 November 2024

1. Introduction

The kidneys are a filtration system for removing waste and contaminants from the body through the urine. The kidneys remove excess acid from the body and maintain healthy levels of salts, water, and minerals in our blood, which are necessary for healthy tissue function. They also produce hormones that assist in blood pressure regulation, the creation of red blood cells, and maintaining healthy bones [1]. Chronic kidney disease (CKD) is a "permanent, gradual, and progressive loss of kidney function over several years caused by a variety of medical conditions" [2]. CKD does not have a known cure, so once it is present, it will last the full life of the individual, and without proper management, it can lead to tissue failure.

CKD is not restricted to a particular group or age. However, patients with diabetes mellitus, heart disease, or high blood pressure are most likely to suffer from CKD. Patients over the age of 60, or with a family history of CKD, or who are HIV positive are at an elevated risk of developing CKD [3]. CKD is a challenge for public health globally, in part due to the preliminary stages of the disease often going undetected, with patients unaware of their condition until it becomes critical or even terminal [4].

The World Health Organization estimates that CKD affects 10% of the global population [5]. CKD has a higher prevalence and morbidity in developing nations due to a lack of access to adequate healthcare among most of the population [5].

1.1. CKD Stages

Chronic kidney disease is categorized into five stages; this helps doctors to determine the best care for each patient. CKD progresses from stage 1 to stage 5; a patient's stage is determined with the estimated glomerular filtration rare (eGFR) [6].

Patients with stage 1 CKD have mild kidney damage with an eGFR of 90 mL/min or greater. At stage 1, there are no symptoms to suggest that the kidneys are damaged as the kidneys appear to function normally at this stage, and most people with the condition are unaware that they have stage 1 CKD [7]. At stage 2, the patient has mild kidney damage with an eGFR of 60–89 mL/min, and as in stage 1 CKD, there are usually no symptoms noticed by the patient, as the kidneys still function normally. At these stages, if the patient discovers that they have CKD, it is usually because they were examined for another medical condition [8].

A patient with stage 3 CKD has moderate to severe kidney damage with an eGFR between 30 and 59 mL/min. This stage is subdivided into stage 3a (mild to moderate kidney damage) with an eGFR of 45–59 mL/min and stage 3b (moderate to severe kidney damage) with an eGFR of 30–44 mL/min. At this stage, the kidneys do not function effectively; as kidney function declines, waste products build up in the blood, resulting in uremia. At stage 3, the patient develops complications such as high blood pressure, anemia, and early bone disease. At stage 3, symptoms may include frequent or less frequent urination, urination changes (foamy; dark orange, brown, tea-colored, or red), fatigue, itchy or dry skin, nausea, loss of appetite that leads to weight loss, and swelling in the hands or feet [9].

A patient with stage 4 CKD has severe loss of kidney function with an eGFR within the range of 15–29 mL/min. In stage 4 CKD, the patient is at the highest risk of having kidney failure and an elevated risk of heart disease. At this stage, the patient will require hemodialysis, peritoneal dialysis, or a kidney transplant. It is characterized as the last stage before kidney failure. At stage 4, damage to the kidney is irreversible. However, under the guidance of a nephrologist and dietitian, efforts may be taken to slow down the kidney damage and mitigate it from progressing to stage 5 (total failure). Unfortunately, even with treatment, kidneys may still fail [10].

Patients with stage 5 kidney failure, commonly known as end-stage kidney disease (ESKD), have an estimated (eGFR) of <15 mL/min. At stage 5, the patient is at the highest risk of developing heart disease. Patients in stages 5 and 4 experience health complications such as anemia, metabolic acidosis, mineral and bone disorders, and hyperkalemia, as well as symptoms such as ammonia-smelling breath, changes in skin color, muscle cramps, difficulty breathing, and little to no urine. Patients diagnosed with stage 5 CKD should consult a nephrologist immediately. At this stage, the patient will need hemodialysis, peritoneal dialysis, or a kidney transplant [11].

Treatment for CKD is expensive and often unsuccessful. Therefore, early detection of CKD is particularly important for implementing preventative measures and managing the disease to slow its progression. Initially, the diagnosis of CKD by a health practitioner involves requesting the patient's medical history, medications they are currently on, and a description of symptoms. Following that, the healthcare practitioner requests a blood test, also known as the estimated glomerular filtration rate (eGFR), and a urine test, also known as the urine albumin–creatinine ratio (uACR), to assess kidney function. Having an eGFR under 60 and a uACR over 30 for three months or more is a sign of kidney disease [12].

1.2. Machine Learning Applied to CKD

Machine learning research in this field has focused on a variety of applications, including predicting the progression of diabetic kidney disease [13], the creation of deep learning models for the prediction of intradialytic hypertension [14], and technology focused on patients with oliguric acute kidney injury [15]. Technologies focused specifically on the prediction of CKD have been developed, using the same public domain dataset relied upon in this study, including a rotation forest model yielding 99.2% accuracy [16], an approach based on the random forest yielding 97% accuracy [17], a support vector machine approach yielding 98.86% accuracy [18], an approach using the K-Nearest Neighbour classifier as well as the Extra Tree Classifier yielding 99% accuracy [19], and multiple approaches based on the XGBoost algorithm (https://github.com/dmlc/xgboost, accessed on 24 February 2024) achieving 94% accuracy [20] and 98.3% accuracy [21]. Additional studies have been conducted on alternative CKD datasets as well [22–24]. It should be noted that there are inevitably differences in the validation employed by each of these studies, as standard validation requires data sample randomization, which would have been performed multiple times independently across studies. As such, direct comparison of these performance metrics has inherent challenges, and coming to firm conclusions as to which algorithms perform the best for prediction of CKD should involve a controlled and fair comparison across all algorithms being considered. Performance metrics reported would ideally be part of an identical validation suite, rather than being compared across studies. Regardless, there is value in comparing machine learning performance reported across studies, even though differences always exist in the evaluation methods employed. It is noteworthy that all studies were able to achieve generally high accuracy results with clear room for further improvement of the technology's predictive capacity.

Although a wide variety of machine learning approaches have been developed for CKD, our study attempted to address the potential value of clinical patient characteristics in this publicly available dataset. In this study, we hypothesized that the use of open-source software df-analyze, inclusive of the high-performing light gradient boosting machine (LGBM), would provide valuable improvements to existing technologies for diagnostics of CKD. We further hypothesized that the inclusion of patient clinical variables, not focused on in other studies based on this dataset, would help further improve the quality of diagnostic predictions in CKD. Finally, we hypothesized that unsupervised learning could assist in the characterization of CKD.

2. Materials and Methods

The dataset was collected from Kaggle (https://www.kaggle.com/datasets/mansoordaku/ckdisease) and is also available from the University of California Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease), and was accessed on 28 January 2024. This dataset is publicly available and published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. This dataset was created from a population of 400 patients over a period of 2 months in India, with the acquisition of 24 feature measurements collected with the target variable CKD status [25]. It has 12 numerical columns and 13 categorical or nominal values. In this dataset, 250 of the samples were from patients with CKD, and 150 samples were acquired from patients without CKD. The machine learning task addressed in this analysis is to predict the CKD status from the set of feature measurements acquired from each patient, thus we are using AI technology to perform CKD diagnoses. To summarize our methods, we downloaded the above dataset and divided it into two parts, one with and one without our patient clinical characteristics, and then public domain software was used to fairly compare a collection of machine learning and feature selection algorithms on each of the two versions of the dataset considered. An unsupervised learning cluster analysis was also performed to help assess sample groupings in the dataset. Details are provided below.

2.1. Dataset Description

Tables 1 and 2 below provide the different variables with their description, data type, and units along with the mean and standard deviation values for the numerical variables for both the CKD and not CKD groups.

Table 1. CKD dataset description for numerical variables.

Variable	Definition	Type	Value	Class of Interest = CKD		Class of Not Interest = NOTCKD	
				Mean	Standard Deviation	Mean	Standard Deviation
age—age	Age of the patient	Numerical	years	54.5	17.4	46.3	15.5
bp—blood pressure (Diastolic)	Pressure of blood pumped by heart into wall's vessels	Numerical	mm/Hg	79.6	15.2	71.4	8.6
sg—specific gravity	Specific indicator of renal function; varies between 1.005 and 1.025 for humans	Numerical	(1.005, 1.010, 1.015, 1.020, 1.025)	1.0	0.0	1.0	0.0
bgr—blood glucose random	Measure of the glucose in the blood at the moment of the test	Numerical	mgs/dL	175.4	92.1	107.7	18.6
bu—blood urea	Measurement of urea nitrogenic blood or serum, a major indicator of kidney failure	Numerical	mgs/dL	72.4	58.6	32.7	11.4
sc—serum creatinine	Creatinine is a waste component that is removed by the kidneys from the blood	Numerical	mgs/dL	4.4	7.0	0.9	0.3
sod—sodium	Concentration of sodium in blood	Numerical	mEq/L	133.9	12.4	141.7	4.8
pot—potassium	Concentration of potassium in blood	Numerical	mEq/L	4.9	4.3	4.3	0.6
hemo—hemoglobin	Hemoglobin is a protein in charge of transport of oxygen in the red blood cells	Numerical	gms	10.6	2.2	15.2	1.3
pcv—packed cell volume	Proportion of blood cells within serum	Numerical	percentage	32.9	7.2	46.3	4.1
wc—white blood cell count	Count of white blood cells	Numerical	cells/cmm	9069.5	3580.5	7687.3	1833.2
rc—red blood cell count	Count of red blood cells	Numerical	millions/cmm	3.9	0.9	5.4	0.6

Table 2. CKD dataset description for categorical variables.

Variable	Definition	Counts for Each Value	Values as a Percentage (%) of All Samples
al—albumin	Measurement of albumin protein in blood	(199:0, 44:1, 43:2, 43:3, 24:4, 1:5, 46:undefined)	(49.75% 0, 11% 1, 10.75% 2, 10.75% 3, 6% 4, 0.25% 5, 11.5% undefined)
su—sugar	Measurement of sugar (glucose) in blood	(290:0, 13:1, 18:2, 14:3, 13:4, 3:5, 49:undefined)	(72.5% 0, 3.25% 1, 4.5% 2, 3.5% 3, 3.25% 4, 0.75% 5, 12.25% undefined)
rbc—red blood cells	Assessment of red blood cells	(201 normal, 47 abnormal, 152 undefined)	(50.25% normal, 11.75% abnormal, 38% undefined)
pc—pus cell	Accumulation of dead white blood cells in the urine	(259 normal, 76 abnormal, 65 undefined)	(64.75% normal, 19% abnormal, 16.25% undefined)
pcc—pus cell clumps	Presence of dead cells in urine, which indicates kidney infection, or a sexually transmitted disease	(42 present, 354 not present, 4 undefined)	(10.5% present, 88.5% not present, 1% undefined)
ba—bacteria	Presence of bacteria in urine	(22 present, 374 not present, 4 undefined)	(5.5% present, 93.5% not present, 1% undefined)

Table 2. *Cont.*

Variable	Definition	Counts for Each Value	Values as a Percentage (%) of All Samples
htn—hypertension	Patient with hypertension diagnosed	(147 yes, 251 no, 2 undefined)	(36.75% yes, 62.75% no, 0.50% undefined)
dm—diabetes mellitus	Failure of the body to react to insulin to control blood glucose levels	(137 yes, 261 no, 2 undefined)	(34.25% yes, 65.25% no, 0.50% undefined)
cad—coronary artery disease	Narrow arteries can cause obstructions of blood flow	(34 yes, 364 no, 2 undefined)	(8.5% yes, 91% no, 0.50% undefined)
appet—appetite	Abnormal appetite	(317 good, 82 poor, 1 undefined)	(79.25% good, 20.5% poor, 0.25% undefined)
pe—pedal edema	Excess fluid in the lower extremities or knees	(76 yes, 323 no, 1 undefined)	(19% yes, 80.75% no, 0.25% undefined)
ane—anemia	Reduction in red blood cells	(60 yes, 339 no, 1 undefined)	(15% yes, 84.75% no, 0.25% undefined)
class—diagnosis	The target value to predict	(250 CKD, 150 not CKD)	(62.5% CKD, 37.5% not CKD)

2.2. Machine Learning

The current framework used for executing the machine learning algorithms considered is the public domain software package df-analyze (https://github.com/stfxecutables/df-analyze/), which is available for download and was accessed on 24 February 2024. "*df-analyze is a command-line tool for performing AutoML on small to medium-sized tabular datasets. In particular, df-analyze attempts to automate:*

1. Feature type inference.
2. Feature description (e.g., univariate associations and stats).
3. Data cleaning (e.g., NaN handling and imputation).
4. Training, validation, and test splitting.
5. Feature selection.
6. Hyperparameter tuning.
7. Model selection and validation.

and saves all key tables and outputs from this process." [26]. This tool was first used in a study on schizophrenia [27] and has since had a variety of features added, including additional feature selection and analytics. Detailed instructions are provided for the use of this software (Instructions).

The software package *df-analyze* was used to fairly compare the following machine learning algorithms: K-Nearest Neighbour (KNN), light gradient boosting machine (LGBM), random forest, logistic regression, Stochastic Gradient Descent, Multi-layered Perceptron Artificial Neural Network, and the support vector machine. A dummy classifier that simply predicts the majority class was also included as a baseline for comparison. All algorithms were compared with two forms of validation: K-fold validation, and evaluation on a holdout dataset. Missing values in the dataset were imputed with the corresponding median value for the numerical attributes after an initial robust normalization scheme documented online, and missing values for categorical variables were converted to a single additional category, which was then clearly labeled as a NaN (Not a Number) indicator after the categorical variable was one-hot encoded (i.e., a new feature is created that is 1 for all samples missing that entry in their categorical variable, and 0 for all other samples). Entries in the dataset containing "?" symbols were converted to NaN, which is compatible with df-analyze alongside blank entries. Hyperparamater tuning was performed within the training set, with the search spaces outlined online for the KNN, LGBM, logistic regression, and Stochastic Gradient Descent. Validation runs were repeated 50 times.

The analysis was repeated two times, once with all the feature measurements available (both numerical features and patient clinical characteristic features), and the second run was performed while excluding the patient clinical characteristic features, which are further presented below to illustrate their predictive potential. Feature selection was included as part of each run of df-analyze, including wrapper, filter (both prediction and association-based), and embedded methods, each applied to each machine learning algorithm included in the analysis.

2.3. Statistical Analysis

For each machine learning and feature selection algorithm pairing trained for CKD prediction as part of df-analyze, we computed the overall accuracy (ACC), area under the receiver operating characteristic curve (AUROC), F1 score, negative predictive value (NPV), positive predictive value (PPV), as well as the sensitivity (sens) and specificity (spec) values. Comparisons of statistical metrics computed were performed on both the holdout set and 5-fold validation on the holdout set, for both the cases with and without the aforementioned clinical patient characteristic features.

3. Results

3.1. Patient Clinical Characteristics

The patient clinical characteristic variables, included in one of our runs of df-analyze but not the other, are outlined in Figures 1 and 2 below, broken down between CKD and not CKD status. These features include whether the patient has the following symptoms: abnormal appetite (APPET), pedal edema (PE), anemia (ANE), hypertension (HTN), diabetes mellitus (DM), and coronary artery disease (CAD). These figures are provided to illustrate the difference between our two runs of our validation software (with and without these features), in order to assess their potential for informing CKD diagnosis, as they were not considered in this manner in previous analyses on this dataset.

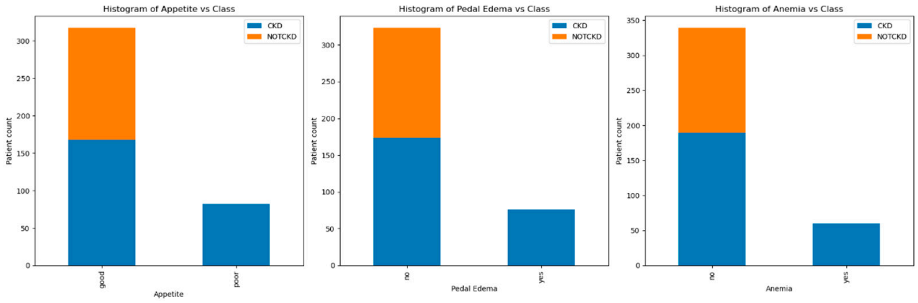


Figure 1. Bar plots of patient clinical characteristics (abnormal appetite, pedal edema, anemia) and how they are distributed in patients with chronic kidney disease (CKD) and patients without (NOTCKD).

It is noteworthy from these bar plots that 100% of the patients who do not have CKD also do not exhibit any of these symptoms in this dataset. Thus, these variables have predictive potential to inform multivariable machine learning technologies, which may contribute towards making more accurate predictions of CKD vs. not CKD.

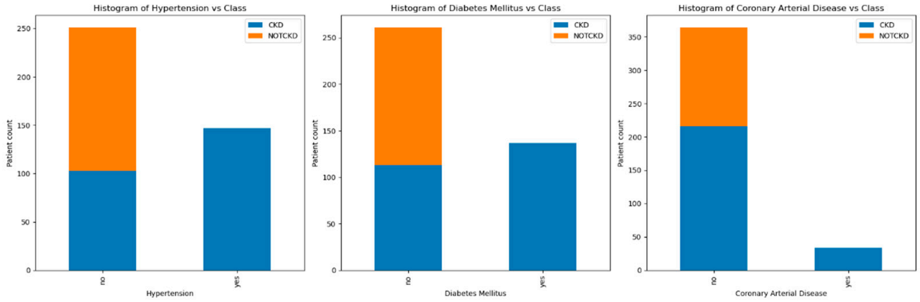


Figure 2. Bar plots of patient characteristics (hypertension—HTN, diabetes mellitus—DM, and coronary artery disease—CAD) and how they are distributed in patients with chronic kidney disease (CKD) and patients without (NOTCKD).

3.2. Machine Learning with Patient Clinical Characteristics

Table 3 provides the results of the machine learning analysis for the complete dataset inclusive of patient clinical characteristics applied to the holdout set. Table 4 provides the machine learning results for the complete dataset inclusive of patient clinical characteristics as part of 5-fold cross-validation. Tables 3 and 4 are provided to illustrate the results of our detailed comparative analysis across machine learning technologies for the entire dataset. The LGBM (light gradient boosting machine) with either no feature selection or with embedded (embed_lgbm)-based feature selection performed the best among all the machine learning techniques compared, producing 99.4 to 100% accuracy (inclusive of clinical patient characteristic features) across the two validation approaches employed. The feature selection reports for the embedded feature selection (embed_lgbm) are provided in Appendix A. The next best performing approach, using prediction-based feature selection (pred), is provided for comparative reference in Appendix B.

Table 3. Holdout set performance for the complete dataset.

Model	Selection	Embed_Selector	ACC	AUROC	BAL-ACC	F1	NPV	PPV	Sens	Spec
lgbm	none	none	1	1	1	1	1	1	1	1
lgbm	embed_lgbm	lgbm	1	1	1	1	1	1	1	1
lgbm	embed_linear	linear	0.994	1	0.995	0.993	0.984	1	0.995	1
lgbm	assoc	none	0.994	1	0.995	0.993	0.984	1	0.995	1
rf	assoc	none	0.994	1	0.992	0.993	1	0.99	0.992	0.983
lr	wrap	none	0.988	0.998	0.99	0.987	0.968	1	0.99	1
rf	embed_lgbm	lgbm	0.988	0.999	0.987	0.987	0.983	0.99	0.987	0.983
lgbm	pred	none	0.988	1	0.99	0.987	0.968	1	0.99	1
mlp	pred	none	0.988	1	0.99	0.987	0.968	1	0.99	1
sgd	wrap	none	0.988	0.998	0.99	0.987	0.968	1	0.99	1
mlp	wrap	none	0.981	0.999	0.985	0.98	0.952	1	0.985	1
rf	pred	none	0.981	0.999	0.978	0.98	0.983	0.98	0.978	0.967
lr	pred	none	0.981	0.998	0.982	0.98	0.967	0.99	0.982	0.983
rf	none	none	0.969	0.997	0.962	0.966	0.982	0.961	0.962	0.933
knn	pred	none	0.963	0.97	0.97	0.961	0.909	1	0.97	1
lgbm	wrap	none	0.956	0.996	0.958	0.954	0.921	0.979	0.958	0.967
rf	embed_linear	linear	0.956	0.992	0.962	0.954	0.908	0.989	0.962	0.983
rf	wrap	none	0.938	0.989	0.937	0.934	0.903	0.959	0.937	0.933
sgd	pred	none	0.906	0.888	0.888	0.897	0.925	0.897	0.888	0.817
knn	assoc	none	0.863	0.927	0.86	0.855	0.797	0.906	0.86	0.85
knn	embed_linear	linear	0.844	0.835	0.835	0.834	0.787	0.879	0.835	0.8
knn	embed_lgbm	lgbm	0.838	0.911	0.83	0.828	0.774	0.878	0.83	0.8
mlp	embed_linear	linear	0.812	0.919	0.773	0.786	0.841	0.802	0.773	0.617
knn	none	none	0.775	0.77	0.77	0.764	0.682	0.84	0.77	0.75
knn	wrap	none	0.769	0.854	0.785	0.765	0.646	0.889	0.785	0.85
mlp	embed_lgbm	lgbm	0.769	0.882	0.778	0.763	0.653	0.871	0.778	0.817
sgd	embed_linear	linear	0.688	0.67	0.67	0.669	0.581	0.755	0.67	0.6
sgd	assoc	none	0.681	0.737	0.648	0.651	0.585	0.729	0.648	0.517
lr	none	none	0.675	0.83	0.583	0.559	0.722	0.669	0.583	0.217
sgd	none	none	0.662	0.654	0.627	0.629	0.558	0.713	0.627	0.483
lr	embed_linear	linear	0.656	0.821	0.548	0.492	0.778	0.649	0.548	0.117
lr	assoc	none	0.644	0.819	0.532	0.462	0.714	0.641	0.532	0.083
lr	embed_lgbm	lgbm	0.631	0.761	0.515	0.43	0.6	0.632	0.515	0.05
mlp	none	none	0.625	0.61	0.5	0.385	nan	0.625	0.5	0
mlp	assoc	none	0.625	0.66	0.5	0.385	nan	0.625	0.5	0
dummy	embed_lgbm	lgbm	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	wrap	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	pred	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
sgd	embed_lgbm	lgbm	0.625	0.593	0.593	0.594	0.5	0.692	0.593	0.467
dummy	none	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	embed_linear	linear	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	assoc	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0

Table 4. Five-fold holdout set performance for complete dataset.

Model	Selection	Embed_Selector	ACC	AUROC	BAL-ACC	F1	NPV	PPV	Sens	Spec
lgbm	pred	none	1	1	1	1	1	1	1	1
lgbm	embed_lgbm	lgbm	0.994	1	0.992	0.993	1	0.99	0.992	0.983
lgbm	assoc	none	0.994	1	0.992	0.993	1	0.99	0.992	0.983
lgbm	none	none	0.994	1	0.992	0.993	1	0.99	0.992	0.983
lgbm	embed_linear	linear	0.994	1	0.992	0.993	1	0.99	0.992	0.983
sgd	wrap	none	0.988	1	0.99	0.987	0.971	1	0.99	1
lgbm	wrap	none	0.981	0.997	0.982	0.98	0.969	0.99	0.982	0.983
rf	embed_linear	linear	0.981	1	0.985	0.98	0.954	1	0.985	1
lr	wrap	none	0.981	1	0.982	0.98	0.971	0.99	0.982	0.983
lr	pred	none	0.975	0.995	0.977	0.974	0.955	0.99	0.977	0.983
rf	wrap	none	0.969	0.994	0.972	0.967	0.937	0.99	0.972	0.983
rf	none	none	0.963	0.995	0.957	0.96	0.966	0.962	0.957	0.933
rf	assoc	none	0.956	0.998	0.962	0.954	0.913	0.99	0.962	0.983
rf	embed_lgbm	lgbm	0.95	0.993	0.957	0.948	0.903	0.99	0.957	0.983
knn	pred	none	0.95	0.957	0.957	0.948	0.903	0.99	0.957	0.983
rf	pred	none	0.944	0.993	0.948	0.941	0.901	0.98	0.948	0.967
sgd	pred	none	0.9	0.897	0.897	0.894	0.864	0.932	0.897	0.883
knn	embed_lgbm	lgbm	0.838	0.91	0.833	0.83	0.784	0.883	0.833	0.817
sgd	embed_linear	linear	0.819	0.815	0.815	0.806	0.775	0.885	0.815	0.8
sgd	assoc	none	0.812	0.889	0.807	0.801	0.744	0.873	0.807	0.783
sgd	none	none	0.812	0.872	0.793	0.797	0.776	0.837	0.793	0.717
knn	assoc	none	0.812	0.929	0.833	0.809	0.693	0.939	0.833	0.917
knn	embed_linear	linear	0.794	0.785	0.785	0.781	0.731	0.849	0.785	0.75
knn	none	none	0.781	0.782	0.782	0.771	0.689	0.862	0.782	0.783
sgd	embed_lgbm	lgbm	0.725	0.715	0.71	0.708	0.648	0.786	0.71	0.65
knn	wrap	none	0.706	0.791	0.738	0.702	0.57	0.905	0.738	0.867
lr	none	none	0.65	0.907	0.533	0.452	1	0.641	0.533	0.067
lr	embed_lgbm	lgbm	0.637	0.855	0.52	0.424	0.75	0.636	0.52	0.05
lr	embed_linear	linear	0.631	0.908	0.508	0.402	1	0.629	0.508	0.017
dummy	embed_lgbm	lgbm	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
lr	assoc	none	0.625	0.907	0.5	0.385	nan	0.625	0.5	0
dummy	wrap	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	pred	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	none	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	embed_linear	linear	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	assoc	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
mlp	assoc	none	0.575	0.642	0.5	0.362	0.375	0.625	0.5	0.2
mlp	pred	none	0.475	0.735	0.5	0.317	0.375	0.625	0.5	0.6
mlp	embed_linear	linear	0.475	0.637	0.5	0.317	0.375	0.625	0.5	0.6
mlp	embed_lgbm	lgbm	0.475	0.63	0.5	0.317	0.375	0.625	0.5	0.6
mlp	wrap	none	0.444	0.724	0.472	0.368	0.35	0.565	0.472	0.583
mlp	none	none	0.375	0.652	0.5	0.273	0.375	nan	0.5	1

3.3. Machine Learning Without Patient Clinical Characteristics

Table 5 provides the results of the machine learning analysis for the reduced dataset, not inclusive of patient clinical characteristics, applied to the holdout set. Table 6 provides the machine learning results for the reduced dataset not inclusive of patient clinical characteristics as part of 5-fold cross-validation. Tables 5 and 6 are provided to illustrate the results of our detailed comparative analysis across machine learning technologies for the reduced dataset that does not include the clinical variables outlined in Figures 1 and 2. Note that there is no clear winning algorithm across the two validation approaches addressed and that the performance of the classifiers is generally degraded relative to the results from the complete dataset inclusive of patient clinical characteristics provided in Tables 3 and 4.

Our leading performing models were consistently the light gradient boosting machine (LGBM) applied to the complete dataset inclusive of clinical patient characteristics. It is noteworthy that the predictive performance of the LGBM with no feature selection was nearly the same as the performance of the LGBM with embedded feature selection (embed_lgbm). The features selected for from the embedded feature selection (embed_lgbm) are summarized in Appendix A, while the features selected for from the filter-prediction-based method (pred) are provided in Appendix B. These appendices provide an indication, based on competing methods, of what feature measurements are most important to rely upon for accurate prediction of CKD.

Table 5. Holdout set performance for dataset without clinical patient characteristics.

Model	Selection	Embed_Selector	ACC	AUROC	BAL-ACC	F1	NPV	PPV	Sens	Spec
lgbm	none	none	1	1	1	1	1	1	1	1
rf	wrap	none	0.994	1	0.995	0.993	0.984	1	0.995	1
lgbm	wrap	none	0.994	1	0.995	0.993	0.984	1	0.995	1
lgbm	pred	none	0.994	1	0.995	0.993	0.984	1	0.995	1
lgbm	assoc	none	0.994	1	0.995	0.993	0.984	1	0.995	1
knn	wrap	none	0.988	0.999	0.99	0.987	0.968	1	0.99	1
mlp	pred	none	0.988	1	0.987	0.987	0.983	0.99	0.987	0.983
lr	wrap	none	0.988	0.999	0.987	0.987	0.983	0.99	0.987	0.983
rf	assoc	none	0.988	1	0.987	0.987	0.983	0.99	0.987	0.983
lgbm	embed_lgbm	lgbm	0.988	1	0.99	0.987	0.968	1	0.99	1
sgd	wrap	none	0.988	0.999	0.987	0.987	0.983	0.99	0.987	0.983
mlp	wrap	none	0.988	1	0.99	0.987	0.968	1	0.99	1
rf	pred	none	0.988	0.998	0.99	0.987	0.968	1	0.99	1
lr	pred	none	0.981	0.999	0.978	0.98	0.983	0.98	0.978	0.967
rf	none	none	0.981	0.999	0.982	0.98	0.967	0.99	0.982	0.983
knn	pred	none	0.975	0.977	0.977	0.974	0.952	0.99	0.977	0.983
lgbm	embed_linear	linear	0.963	0.988	0.957	0.96	0.966	0.961	0.957	0.933
rf	embed_lgbm	lgbm	0.956	0.993	0.952	0.953	0.949	0.96	0.952	0.933
rf	embed_linear	linear	0.944	0.988	0.938	0.94	0.932	0.95	0.938	0.917
lr	embed_lgbm	lgbm	0.925	0.98	0.917	0.919	0.914	0.931	0.917	0.883
sgd	pred	none	0.912	0.935	0.907	0.907	0.883	0.93	0.907	0.883
sgd	embed_lgbm	lgbm	0.906	0.969	0.892	0.898	0.909	0.905	0.892	0.833
knn	embed_linear	linear	0.869	0.932	0.855	0.859	0.842	0.883	0.855	0.8
knn	embed_lgbm	lgbm	0.85	0.92	0.847	0.842	0.781	0.896	0.847	0.833
knn	assoc	none	0.812	0.797	0.797	0.799	0.759	0.843	0.797	0.733
mlp	assoc	none	0.8	0.885	0.747	0.762	0.889	0.774	0.747	0.533
knn	none	none	0.787	0.877	0.783	0.777	0.697	0.851	0.783	0.767
mlp	embed_lgbm	lgbm	0.775	0.97	0.7	0.709	1	0.735	0.7	0.4
mlp	none	none	0.775	0.896	0.783	0.769	0.662	0.872	0.783	0.817
mlp	embed_linear	linear	0.762	0.852	0.777	0.758	0.641	0.878	0.777	0.833
lr	none	none	0.738	0.865	0.67	0.675	0.8	0.723	0.67	0.4
lr	assoc	none	0.738	0.865	0.67	0.675	0.8	0.723	0.67	0.4
lr	embed_linear	linear	0.719	0.866	0.645	0.645	0.778	0.707	0.645	0.35
sgd	assoc	none	0.681	0.632	0.632	0.635	0.605	0.709	0.632	0.433
sgd	embed_linear	linear	0.681	0.635	0.635	0.639	0.6	0.713	0.635	0.45
sgd	none	none	0.662	0.79	0.603	0.603	0.579	0.689	0.603	0.367
dummy	embed_lgbm	lgbm	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	wrap	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	pred	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	none	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	embed_linear	linear	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	assoc	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0

Table 6. Five-fold holdout set performance for dataset without clinical patient characteristic variables.

Model	Selection	Embed_Selector	ACC	AUROC	BAL-ACC	F1	NPV	PPV	Sens	Spec
sgd	wrap	none	0.981	1	0.982	0.98	0.969	0.99	0.982	0.983
lr	wrap	none	0.981	1	0.982	0.98	0.969	0.99	0.982	0.983
knn	wrap	none	0.975	0.995	0.98	0.974	0.941	1	0.98	1
lgbm	none	none	0.975	0.998	0.97	0.973	0.985	0.972	0.97	0.95
lgbm	assoc	none	0.975	0.997	0.97	0.973	0.985	0.972	0.97	0.95
lgbm	embed_lgbm	lgbm	0.975	0.998	0.97	0.973	0.985	0.972	0.97	0.95
lgbm	pred	none	0.969	0.997	0.965	0.966	0.966	0.971	0.965	0.95
lgbm	wrap	none	0.963	0.996	0.963	0.96	0.938	0.981	0.963	0.967
rf	assoc	none	0.963	0.997	0.96	0.96	0.95	0.971	0.96	0.95
rf	none	none	0.956	0.996	0.958	0.954	0.922	0.981	0.958	0.967
lr	pred	none	0.956	0.997	0.958	0.954	0.922	0.981	0.958	0.967
lgbm	embed_linear	linear	0.956	0.992	0.955	0.953	0.938	0.971	0.955	0.95
rf	wrap	none	0.944	0.99	0.942	0.94	0.92	0.96	0.942	0.933

Table 6. *Cont.*

Model	Selection	Embed_Selector	ACC	AUROC	BAL-ACC	F1	NPV	PPV	Sens	Spec
knn	pred	none	0.938	0.94	0.94	0.934	0.895	0.969	0.94	0.95
rf	pred	none	0.931	0.989	0.928	0.927	0.909	0.951	0.928	0.917
rf	embed_linear	linear	0.925	0.982	0.927	0.921	0.881	0.959	0.927	0.933
rf	embed_lgbm	lgbm	0.919	0.992	0.928	0.915	0.846	0.981	0.928	0.967
lr	embed_lgbm	lgbm	0.906	0.973	0.902	0.9	0.873	0.93	0.902	0.883
sgd	pred	none	0.9	0.922	0.893	0.894	0.882	0.92	0.893	0.867
sgd	embed_lgbm	lgbm	0.881	0.955	0.872	0.873	0.852	0.903	0.872	0.833
knn	none	none	0.844	0.945	0.872	0.842	0.73	0.988	0.872	0.983
knn	embed_linear	linear	0.844	0.95	0.838	0.834	0.789	0.89	0.838	0.817
knn	embed_lgbm	lgbm	0.831	0.924	0.835	0.825	0.765	0.903	0.835	0.85
sgd	none	none	0.806	0.875	0.795	0.794	0.746	0.85	0.795	0.75
sgd	assoc	none	0.794	0.785	0.785	0.782	0.725	0.845	0.785	0.75
knn	assoc	none	0.787	0.793	0.793	0.78	0.687	0.879	0.793	0.817
sgd	embed_linear	linear	0.781	0.762	0.762	0.764	0.728	0.815	0.762	0.683
lr	none	none	0.631	0.876	0.512	0.415	0.75	0.631	0.512	0.033
lr	assoc	none	0.631	0.876	0.512	0.415	0.75	0.631	0.512	0.033
dummy	embed_lgbm	lgbm	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	wrap	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	pred	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	none	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	embed_linear	linear	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
dummy	assoc	none	0.625	0.5	0.5	0.385	nan	0.625	0.5	0
lr	embed_linear	linear	0.619	0.872	0.495	0.382	0	0.623	0.495	0
mlp	embed_lgbm	lgbm	0.556	0.724	0.525	0.402	0.292	0.709	0.525	0.4
mlp	pred	none	0.537	0.813	0.543	0.4	0.435	0.702	0.543	0.567
mlp	wrap	none	0.525	0.736	0.5	0.34	0.375	0.625	0.5	0.4
mlp	none	none	0.475	0.619	0.5	0.317	0.375	0.625	0.5	0.6
mlp	embed_linear	linear	0.475	0.579	0.5	0.317	0.375	0.625	0.5	0.6
mlp	assoc	none	0.375	0.633	0.5	0.273	0.375	nan	0.5	1

3.4. Cluster Analysis Results

In order to help illustrate and improve understanding of the complete dataset, we also performed a cluster analysis with the unsupervised K-means algorithm. For the purpose of determining an optimal K value, both silhouette scores and Calinski–Harabasz scores [28] were applied, respectively, producing Figures 3 and 4.

Based on both the silhouette score and the Calinski–Harabasz score, the highest K value occurred with K = 2. We were also separately interested in K = 6, as this value exhibited a remarkable drop-off in the silhouette score when K was further raised beyond 6, indicating possible value from this granularity of clustering. Another reason why we were interested in the K value of 6 is that there are five distinct stages of CKD, as outlined in the introduction, as well as a normal/healthy control group (6 total). The clustering results with K = 2 are visualized using a principal component analysis (PCA) projection of the data in Figure 5, showing loose alignment with the main group status (CKD vs. not CKD). Furthermore, the PCA-projected clustering with K = 6 produced the results shown in Figure 6, which appear to loosely align with the severity of CKD, implying that the feature measurements included in this analysis hold considerable value, a finding with relevance for possible future applications of CKD staging. This implies the potential for extending machine learning technology in this domain, to be used in stage predictions and monitoring of disease progression. Figure 7 is provided as a comparative reference for Figures 5 and 6, demonstrating the ground truth diagnoses of CKD and not CKD. This is further reinforced by the resultant cluster centers, with centroids illustrating characteristic underlying feature profiles, the results of which are provided in Figure 8, which demonstrates group-wise differences between the learned clusters.

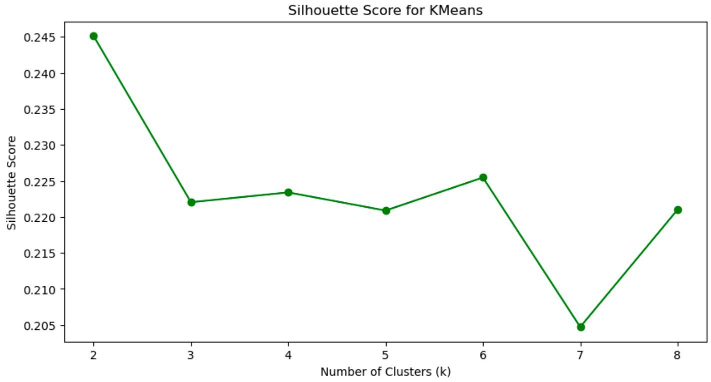


Figure 3. Silhouette score applied to the dataset. This plot was used to assess the apparent value of clustering results at varying numbers of clusters. The highest values and local peaks were of potential interest.

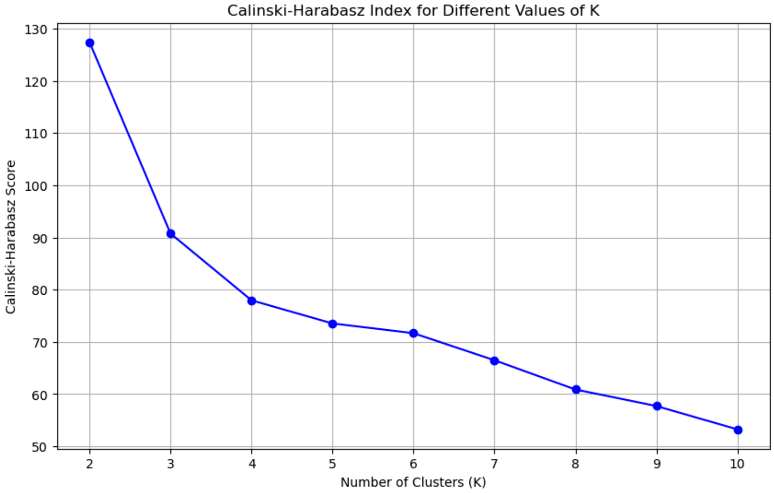


Figure 4. Calinski–Harabasz score plot for the same dataset. This plot was used to assess the apparent value of clustering results at varying numbers of clusters. The highest values and local peaks were of potential interest.

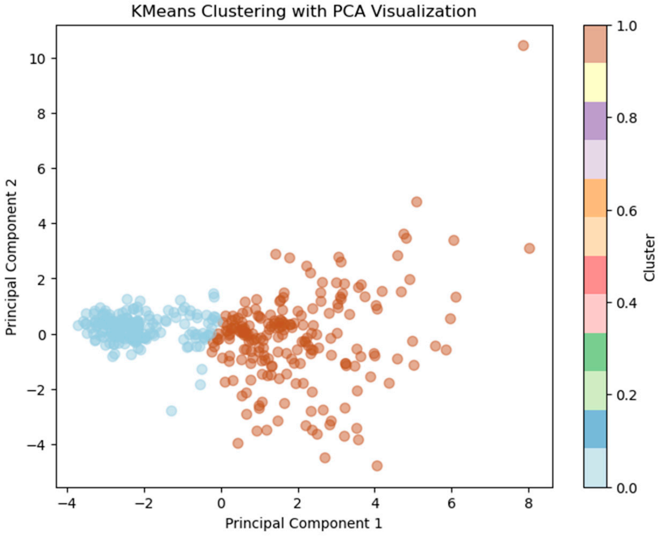


Figure 5. Plot of resulting clusters, using K-means algorithm with PCA and K = 2.

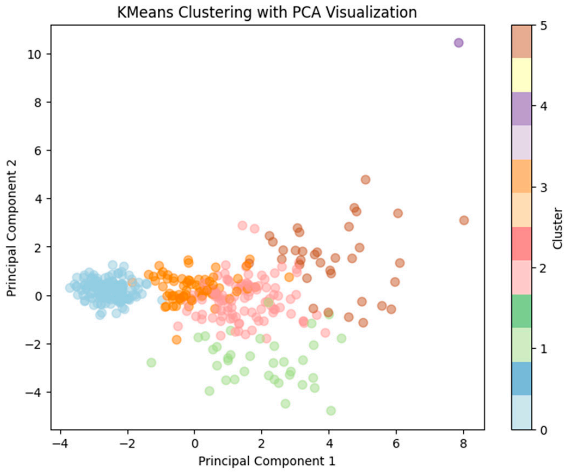


Figure 6. Plot of resulting clusters, using K-means algorithm with PCA and K = 6.

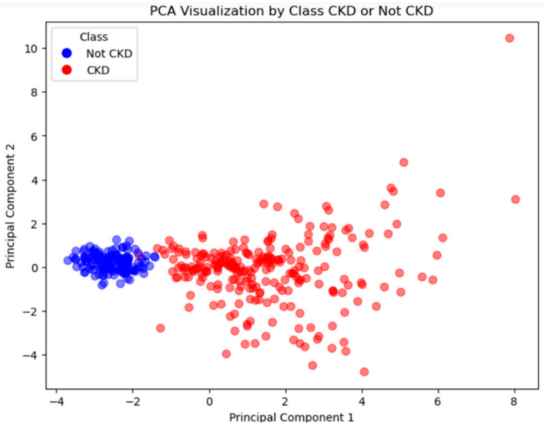
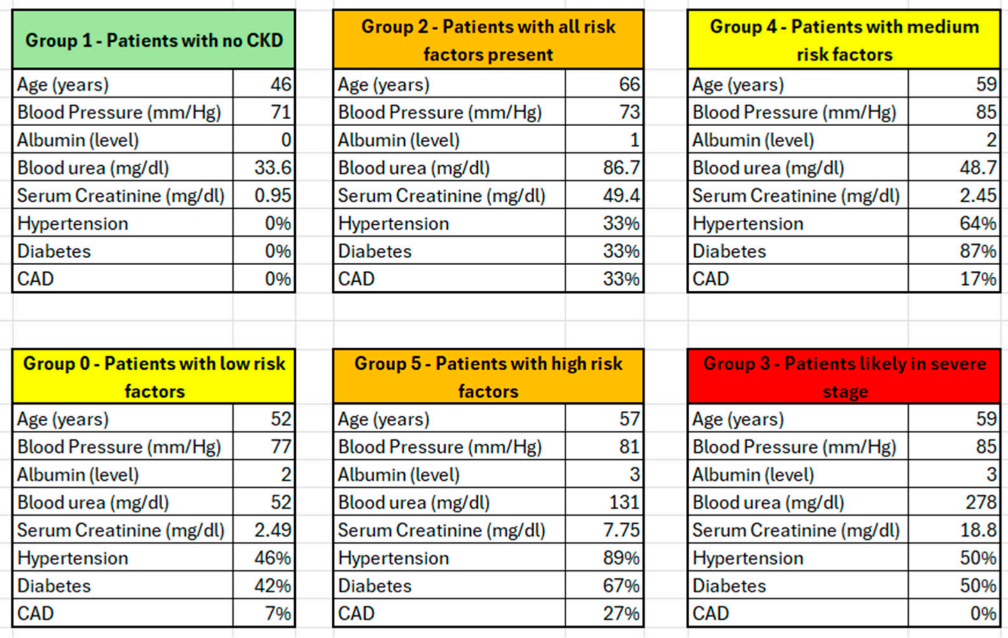


Figure 7. Visualization of PCA components' projection based on the class CKD or not CKD.



Group 1 - Patients with no CKD	
Age (years)	46
Blood Pressure (mm/Hg)	71
Albumin (level)	0
Blood urea (mg/dl)	33.6
Serum Creatinine (mg/dl)	0.95
Hypertension	0%
Diabetes	0%
CAD	0%

Group 2 - Patients with all risk factors present	
Age (years)	66
Blood Pressure (mm/Hg)	73
Albumin (level)	1
Blood urea (mg/dl)	86.7
Serum Creatinine (mg/dl)	49.4
Hypertension	33%
Diabetes	33%
CAD	33%

Group 4 - Patients with medium risk factors	
Age (years)	59
Blood Pressure (mm/Hg)	85
Albumin (level)	2
Blood urea (mg/dl)	48.7
Serum Creatinine (mg/dl)	2.45
Hypertension	64%
Diabetes	87%
CAD	17%

Group 0 - Patients with low risk factors	
Age (years)	52
Blood Pressure (mm/Hg)	77
Albumin (level)	2
Blood urea (mg/dl)	52
Serum Creatinine (mg/dl)	2.49
Hypertension	46%
Diabetes	42%
CAD	7%

Group 5 - Patients with high risk factors	
Age (years)	57
Blood Pressure (mm/Hg)	81
Albumin (level)	3
Blood urea (mg/dl)	131
Serum Creatinine (mg/dl)	7.75
Hypertension	89%
Diabetes	67%
CAD	27%

Group 3 - Patients likely in severe stage	
Age (years)	59
Blood Pressure (mm/Hg)	85
Albumin (level)	3
Blood urea (mg/dl)	278
Serum Creatinine (mg/dl)	18.8
Hypertension	50%
Diabetes	50%
CAD	0%

Figure 8. Centroids of each cluster from K-means analysis of the CKD dataset, K = 6. Note that group numbers are arbitrarily assigned by the K-means algorithm.

3.5. Results Summary

In summary, our findings demonstrate that the inclusion of patient clinical characteristics is associated with a broad improvement in predictive accuracy across algorithms when compared to an analysis where they were excluded, an approach not considered in pre-existing studies. Our findings also indicate that multiple configurations of the LGBM consistently produce high-performing models with accuracies ranging from 99.4 to 100%. Cluster analysis results demonstrate the potential for machine-learning-based characterization of CKD severity, with potential long-term applications in CKD staging.

4. Discussion

4.1. Patient Clinical Characteristics

It is noteworthy from Figures 1 and 2 that non CKD patients are distributed in just one of the two possible values for these clinical variables. Thus, these clinical patient characteristic variables have predictive potential in informing multivariable machine learning technologies, which would allow them to contribute towards making more accurate predictions of CKD vs. not CKD.

4.2. Impact of Patient Clinical Characteristics on Machine Learning

Our results indicate that the approach developed achieved 99.4 to 100% accuracy across multiple types of validation. When applying the same analyses but on the reduced version of the dataset, not including the set of clinical patient characteristics, the performance degrades to be in line with the results presented in previous literature studies, outlined in the introduction. These results imply value from the inclusion of clinical patient characteristics in the prediction/diagnosis of CKD by machine learning. This information could be of potential value to clinicians involved in the management of patients with CKD. Our leading feature-reduced models, based on LGBM and embedded feature selection, indicate useful predictive value from knowledge of the patient's diabetes mellitus status (see Appendix A). Our findings also demonstrate that an unsupervised machine learning analysis, inclusive of patient clinical characteristics, loosely aligns with the severity of CKD. These findings lead us to propose that future work should involve the development of CKD characterization/staging technologies and potentially disease progression and treatment monitoring technologies, which may be of future clinical interest.

4.3. Comparative Machine Learning Results Across Studies

The leading models were obtained with the light gradient boosting machine (LGBM), noteworthy in that it is a lightweight computationally efficient alternative to the extreme gradient boosting machine, which was the leading machine learning technique in several studies on this topic [20,21,24]. High-quality results were obtainable with the LGBM inclusive of all of our features, but we were also able to produce similarly performing models with reduced feature sets, outlined in Appendices A and B. These findings imply that feature measurements not selected for inclusion in Appendices A and B are either not useful for CKD diagnosis by machine learning, or are redundant to other features included in the selection process. Our results compare favorably with existing research focused on the same dataset, including a rotation forest model yielding 99.2% accuracy [16], an approach based on the random forest yielding 97% accuracy [17], a support vector machine approach yielding 98.86% accuracy [18], an approach using the K-Nearest Neighbour classifier as well as the Extra Tree Classifier yielding 99% accuracy [19], and multiple approaches based on the XGBoost algorithm (https://github.com/dmlc/xgboost, accessed on 24 February 2024) achieving 94% accuracy [20], and 98.3% accuracy [21]. Thus, our finding that multiple configurations of the LGBM can produce 99.4 to 100% accuracy across two validation methods potentially adds value to the machine learning literature on this topic. It should also be noted that the LGBM is a computationally efficient 'lightweight' machine learning algorithm, so it offers advantages in terms of computational complexity.

4.4. Limitations and Future Work

Limitations of this study include that it was performed on a dataset with relatively few samples (*n* = 400). Machine learning algorithms often achieve peak performance on balanced datasets (i.e., in the context of this work, a balanced dataset would have an equal number of samples in the group of interest—CKD—and in the control group). Our dataset is not balanced, and as such, this can cause a potential bias, possibly reducing the predictive accuracy of the technologies trained (on unbalanced data). Acquiring a larger, balanced dataset may help reduce associated machine learning class imbalance bias in this application. Noteworthy, however, is that if you sample a large population, people with CKD are expected to be a small minority. As such, a balanced dataset with equal numbers of CKD and controls would not be expected to be representative of the population that such a tool would be tasked with diagnosing in the real world. Training the algorithm on a non-representative population relative to how the tool will be used in the real world can cause a different type of bias, in that the performance estimates from validation on the balanced dataset will necessarily shift to align with the reality of the technology's use on a population in which CKD is comparatively rare. Future work will involve evaluating the approaches developed herein on larger and independently acquired datasets. Future work should also involve a direct comparison of high-performing competing algorithms applied in similar studies on CKD, such as XGBoost [20,21,24] and the Rotation Forest [16]. Strengths of this study include the use of open-source machine learning technology (https://github.com/stfxecutables/df-analyze/), inclusive of a standardized validation suite and an array of feature selection technologies, the consideration of the light gradient boosting algorithm, and the consideration of patient clinical characteristics in the models developed, providing a thorough analysis with high-performing predictive results. The inclusion of an unsupervised clustering analysis is another strength of this study.

Many people live with CKD and are unaware of their status until it reaches an advanced stage of development [4]. As such, the lack of efficient diagnostics applied to asymptomatic populations results in many cases of CKD going untreated, and thus, likely progressing towards more severe manifestations of the condition. In this context, AI technologies that can be easily applied to large populations of individuals, which has a lot of potential to improve the clinical management of CKD by identifying the condition in asymptomatic people through screening and thus supporting earlier treatment, which is commonly associated with better prognoses/outcomes.

The advantages of the approaches taken in this study and in the literature [16–24] include the creation of objective learning machines that are approaching perfect accuracy for the diagnosis of the condition. The major limitations of using artificial intelligence, at this time, include limited datasets and samples upon which to train the learning machines, potentially limiting their robust applicability in the real world. The limited explainability of existing machine learning technologies is another limitation of the use of the current methods, as when an AI technology makes a diagnosis, we want to be able to report the specific logical reasons why the AI came to its diagnostic conclusion. Unfortunately, this is largely an unsolved problem in AI technologies generally.

The work presented herein has the potential to assist with early diagnoses, potentially accurately predicting patient CKD status in the early stages of disease progression, which could theoretically result in earlier interventions and thus an improved standard of patient care and improved prognoses/outcomes. While the population in this study is relatively small (*n* = 400), such high-performing technologies could be the subject of future research in CKD screening, potentially helping to identify the disease prior to patients exhibiting symptoms, which has tremendous potential for reducing the morbidity associated with the condition. Machine learning clustering results demonstrate the potential for further characterizing CKD into stages and creating technologies that can assist in the monitoring of disease and treatment progression. Future work will investigate the staging of CKD with a dataset that contains ground truth knowledge of CKD stages, established by clinical experts in the field, which was unavailable in our dataset.

5. Conclusions

In summary, the progression of chronic kidney disease can be managed if detected early, which can help prevent damage to the kidney, and thus the development of technologies that can assist in early detection has considerable potential for improving the standard of patient care in CKD. This study highlights the efficacy of artificial intelligence and machine learning in diagnosing chronic kidney disease. Using public domain software, we demonstrated that the light gradient boosting machine (LGBM) with embedded feature selection performed the best across all metrics, with accuracies ranging from 99.4 to 100% across multiple validation trials. Cluster analysis results demonstrate potential towards the creation of staging technology for disease characterization, as well as treatment and disease progression technologies.

Author Contributions: Conceptualization, J.F., P.E., A.K.S. and J.L.; methodology, J.F., P.E. and A.K.S.; df-analyze software, D.B. and J.L.; validation, D.B. and J.L.; formal analysis, J.F., P.E., A.K.S. and J.L.; investigation, J.F., P.E. and A.K.S.; resources, D.B. and J.L.; data curation, J.F., P.E. and A.K.S.; writing—original draft preparation, P.E., J.F. and A.K.S.; writing—review and editing, P.E., J.F. and J.L.; visualization, J.F.; supervision, J.L.; project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by a Canada Foundation for Innovation grant, a Nova Scotia Research and Innovation Trust grant, an NSERC Discovery grant, a Compute Canada Resource Allocation, and a Nova Scotia Health Authority grant to J.L.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed from Kaggle at https://www.kaggle.com/datasets/mansoordaku/ckdisease and the University of California Irvine (UCI) Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease (accessed on 24 February 2024). No new data were created or collected specifically for this study. Since this was a retrospective analysis of public domain data, no institutional review board approval was necessary.

Acknowledgments: We would like to extend our heartfelt thanks to St. Francis Xavier University and the Department of Computer Science for providing the computational resources and facilities that made this research possible. We would like to express our sincere gratitude to Jacob Levman for his invaluable guidance and supervision throughout this project. His expertise and insight were instrumental in our research, and his support with both the technical aspects and report preparation was greatly appreciated; without him all of this would not have been possible. We are also grateful to Derek Berger and Jacob Levman for their key role in developing and maintaining the df-analyze software, which was fundamental to our analysis. Finally, we specifically extend our gratitude to all those who offered technical support and guidance throughout the project such as Xuchen Guo.

Conflicts of Interest: Levman is founder of Time Will Tell Technologies, Inc. The authors declare no relevant conflicts of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

The results of embedded feature selection (embed_lgbm) with the LGBM algorithm on the entire dataset.

```
# Wrapper-Based Feature Selection Summary
Wrapper model: LGBM
## Selected Features
['sod_NAN', 'rc_NAN', 'age', 'sg', 'al', 'bgr', 'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc',
'dm_yes', 'pc_nan', 'rbc_normal']
## Selection scores (Importances: Larger magnitude = More important)
| feature | score |
|:------------|----------:|
| age_NAN | 0.000e+00 |
| bp_NAN | 0.000e+00 |
| sg_NAN | 0.000e+00 |
| al_NAN | 0.000e+00 |
| su_NAN | 0.000e+00 |
| bgr_NAN | 0.000e+00 |
| bu_NAN | 0.000e+00 |
| sc_NAN | 0.000e+00 |
| sod_NAN | 1.700e+01 |
| pot_NAN | 1.200e+01 |
| hemo_NAN | 6.000e+00 |
| pcv_NAN | 2.000e+00 |
| wc_NAN | 3.000e+00 |
| rc_NAN | 2.400e+01 |
| age | 2.800e+01 |
| bp | 1.200e+01 |
| sg | 3.800e+01 |
| al | 3.800e+01 |
| su | 0.000e+00 |
| bgr | 2.500e+01 |
| bu | 1.100e+01 |
| sc | 8.000e+01 |
| sod | 2.600e+01 |
| pot | 2.700e+01 |
| hemo | 1.160e+02 |
| pcv | 2.400e+01 |
| wc | 2.000e+01 |
| rc | 1.100e+01 |
| ane_yes | 0.000e+00 |
| ane_nan | 0.000e+00 |
| appet_poor | 5.000e+00 |
| appet_nan | 0.000e+00 |
| ba_present | 0.000e+00 |
| ba_nan | 0.000e+00 |
| cad_yes | 0.000e+00 |
| cad_nan | 0.000e+00 |
| dm_yes | 1.800e+01 |
| dm_nan | 0.000e+00 |
| htn_yes | 1.300e+01 |
| htn_nan | 0.000e+00 |
| pc_normal | 3.000e+00 |
| pc_nan | 1.700e+01 |
| pcc_present | 0.000e+00 |
| pcc_nan | 0.000e+00 |
| pe_yes | 0.000e+00 |
| pe_nan | 0.000e+00 |
| rbc_normal | 1.100e+02 |
| rbc_nan | 4.000e+00 |
```

Appendix B

The results of filter-prediction-based feature selection (pred) with the LGBM algorithm on the entire dataset.

```
# Filter-Based Feature Selection Summary
## Selected Features
['hemo', 'pcv', 'rc', 'sc', 'sg', 'al', 'sod', 'rbc', 'htn', 'dm', 'pc', 'ba']
## Selection Prediction Scores
```

```
### Continuous Features (Accuracy: Higher = More important)
| feature | acc |
|:----------|----------:|
| hemo | 2.750e-01 |
| pcv | 2.583e-01 |
| rc | 2.208e-01 |
| sc | 2.083e-01 |
| sg | 1.417e-01 |
| al | 1.417e-01 |
| sod | 1.333e-01 |
| bgr | 6.667e-02 |
| bp | 3.750e-02 |
| age | 3.333e-02 |
| bu | 2.917e-02 |
| su | 0.000e+00 |
| pot | 0.000e+00 |
| wc | 0.000e+00 |
### Categorical Features (Accuracy: Higher = More important)
| feature | acc |
|:----------|----------:|
| rbc | 2.000e-01 |
| htn | 1.125e-01 |
| dm | 8.333e-02 |
| pc | 2.917e-02 |
| ba | 1.250e-02 |
| pcc | 1.250e-02 |
| ane | 0.000e+00 |
| appet | 0.000e+00 |
| cad | 0.000e+00 |
| pe | 0.000e+00 |
```

References

1. National Institute of Diabetes and Digestive and Kidney Diseases. Your Kidneys & How They Work—National Institute of Diabetes and Digestive and Kidney Diseases. Available online: https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work (accessed on 20 February 2024).
2. Johns Hopkins Hospital. Chronic Kidney Disease | Johns Hopkins Medicine. Available online: https://www.hopkinsmedicine.org/health/conditions-and-diseases/chronic-kidney-disease (accessed on 24 February 2024).
3. National Kidney Foundation. Chronic Kidney Disease (CKD)—Symptoms, Causes, Treatment | National Kidney Foundation. Available online: https://www.kidney.org/atoz/content/about-chronic-kidney-disease (accessed on 24 February 2024).
4. Zhao, J.; Zhang, Y.; Qiu, J.; Zhang, X.; Wei, F.; Feng, J.; Chen, C.; Zhang, K.; Feng, S.; Li, W.D. An early prediction model for chronic kidney disease. *Sci. Rep.* **2022**, *12*, 2765. [CrossRef] [PubMed]
5. World Health Organization. The Top 10 Causes of Death—World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed on 24 February 2024).
6. Estimating Glomerular Filtration Rate. National Institute of Diabetes and Digestive and Kidney Diseases. Available online: https://www.niddk.nih.gov/health-information/professionals/clinical-tools-patient-management/kidney-disease/laboratory-evaluation/glomerular-filtration-rate/estimating (accessed on 3 February 2024).
7. American Kidney Fund. Stage 1 of Chronic Kidney Disease CKD: Causes, Symptoms and Treatment. Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease/stage-1-chronic-kidney-disease (accessed on 24 February 2024).
8. American Kidney Fund. Stage 2 Chronic Kidney Disease (CKD). Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease/stage-2-chronic-kidney-disease-ckd (accessed on 24 February 2024).
9. American Kidney Fund. Stage 3 Chronic Kidney Disease (CKD). Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease/stage-3-chronic-kidney-disease-ckd (accessed on 24 February 2024).
10. American Kidney Fund. Stage 4 Chronic Kidney Disease (CKD). Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease/stage-4-chronic-kidney-disease-ckd (accessed on 24 February 2024).
11. American Kidney Fund. Stage 5 Chronic Kidney Disease (CKD). Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease/stage-5-chronic-kidney-disease-ckd (accessed on 24 February 2024).

12. Cleveland Clinic. Chronic Kidney Disease (CKD): Symptoms & Treatment. Available online: https://my.clevelandclinic.org/health/diseases/15096-chronic-kidney-disease (accessed on 7 May 2023).
13. Chen, T.K.; Knicely, D.H.; Grams, M.E. Chronic Kidney Disease Diagnosis and Management A Review. National Library of Medicine. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7015670 (accessed on 11 March 2024).
14. Lee, H.; Yun, D.; Yoo, J.; Yoo, K.; Kim, Y.C.; Kim, D.K.; Oh, K.H.; Joo, K.W.; Kim, Y.S.; Kwak, N.; et al. Deep learning model for real-time prediction of intradialytic hypotension. *Clin. J. Am. Soc. Nephrol.* **2021**, *16*, 396–406. [CrossRef] [PubMed]
15. Zhang, Z.; Ho, K.M.; Hong, Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit. Care* **2019**, *23*, 112. [CrossRef] [PubMed]
16. Dritsas, E.; Trigka, M. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. *Big Data Cogn. Comput.* **2022**, *6*, 98. [CrossRef]
17. Yashfi, S.; Islam, M.; Pritilata; Sakib, N.; Islam, T.; Shahbaaz, M.; Pantho, S. Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020. [CrossRef]
18. Chittora, P.; Sandeep, C.; Chakrabarti, P.; Kumawat, G.; Chakrabarti, T.; Leonowicz, Z.; Jasiński, M.; Jasiński, Ł.; Gono, R.; Jasińska, E.; et al. Prediction of Chronic Kidney Disease—A Machine Learning Perspective. *IEEE Access* **2021**, *9*, 17312–17334. [CrossRef]
19. Baidya, D.; Umaima, U.; Islam, M.N.; Shamrat, F.M.J.M.; Pramanik, A.; Rahman, M.S. A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1305–1310. [CrossRef]
20. Abdur, M.; Rahat, R.; Cao, D.M.; Tayaba, M.; Ghosh, B.P.; Ayon, H.; Nobe Nur Akter, T.; Rahman, M.; Bhuiyan, M.S. Comparing Machine Learning Techniques for Detecting Chronic Kidney Disease in Early Stage. *J. Comput. Sci. Technol. Stud.* **2024**, *6*, 20–32. [CrossRef]
21. Islam, M.A.; Majumder, M.Z.H.; Hussein, M.A. Chronic kidney disease prediction based on machine learning algorithms. *J. Pathol. Inform.* **2023**, *14*, 100189. [CrossRef] [PubMed]
22. Debal, D.A.; Sitote, T.M. Chronic kidney disease prediction using machine learning techniques. *J. Big Data* **2022**, *9*, 109. [CrossRef]
23. Ghosh, S.K.; Khandoker, A.H. A machine learning driven nomogram for predicting chronic kidney disease stages 3–5. *Sci. Rep.* **2023**, *13*, 21613. [CrossRef] [PubMed]
24. Xiao, J.; Ding, R.; Xu, X.; Guan, H.; Feng, X.; Sun, T.; Zhu, S.; Ye, Z. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J. Transl. Med.* **2019**, *17*, 119. [CrossRef] [PubMed] [PubMed Central]
25. Rubini, L.; Soundarapandian, P.; Eswaran, P. *Chronic Kidney Disease*; UCI Machine Learning Repository: Espoo, Finland, 2015. [CrossRef]
26. DM-Berger. *stfxecutables/df-analyze: Sensitivity/Specificity Update*; Zenodo: Geneva, Switzerland, 2021. [CrossRef]
27. Levman, J.; Jennings, M.; Rouse, E.; Berger, D.; Kabaria, P.; Nangaku, M.; Gondra, I.; Takahashi, E. A morphological study of schizophrenia with Magnetic Resonance Imaging, advanced analytics, and Machine Learning. *Front. Neurosci.* **2022**, *16*, 926426. [CrossRef] [PubMed]
28. Wella, Y.; Okfalisa, O.; Insani, F.; Saeed, F.; Hussin, A. Service quality dealer identification: The optimization of K-Means clustering. *SINERGI* **2023**, *27*, 433. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.