Chronic Kidney Disease Prediction Using Machine Learning Methods

Imesh Udara Ekanayake
Dept. Computer Engineering
University of Peradeniya
Peradeniya, 20400, Sri Lanka
imeshuek@eng.pdn.ac.lk

Damayanthi Herath
Dept. Computer Engineering
University of Peradeniya
Peradeniya, 20400, Sri Lanka
damayanthiherath@eng.pdn.ac.lk

***Abstract*—Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for early prediction of CKD. Machine learning methods are effective in CKD prediction. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. The research also considers the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.**

***Index Terms*—chronic kidney disease, chronic renal disease, machine learning, classification algorithms, extra tree classifier, random forest classifier**

I. Introduction

Kidneys are two bean-shaped organs, each about the size of a fist [1]–[3]. They are located just below the rib cage, one on each side of the spine. Every day, the kidneys filter about 120 to 150 quarts of blood to produce about 1 to 2 quarts of urine. The key function of the kidneys is to remove waste products and excess fluid from the body through the urine. The production of urine involves highly complex steps of excretion and re-absorption. This process is necessary to maintain a stable balance of body chemicals. The critical regulation of the body 's salt, potassium and acid content is performed by the kidneys and produce hormones that affect the function of other organs. For example, a hormone produced by the kidneys stimulates red blood cell production, regulate blood pressure and control calcium metabolism etc.

Chronic kidney disease (CKD) is a major issue worldwide which is a condition characterized by a gradual loss of kidney function over time, 14% of the world population suffer from CKD. Over 2 million people worldwide currently receive treatment with dialysis or a kidney transplant to stay alive, yet this number may only represent 10% of people who need treatment to live. Chronic kidney disease causes more deaths than breast cancer or prostate cancer [2]. The stages of CKD are mainly based on the measured or estimated glomerular filtration rate (eGFR) which is based on creatinine level [4], gender, race and age. There are five stages of kidney functionality [5]. The function is normal in stage 1 and minimally reduced in stage 2 but the majority of cases are at stage 3 (see Fig. 1).

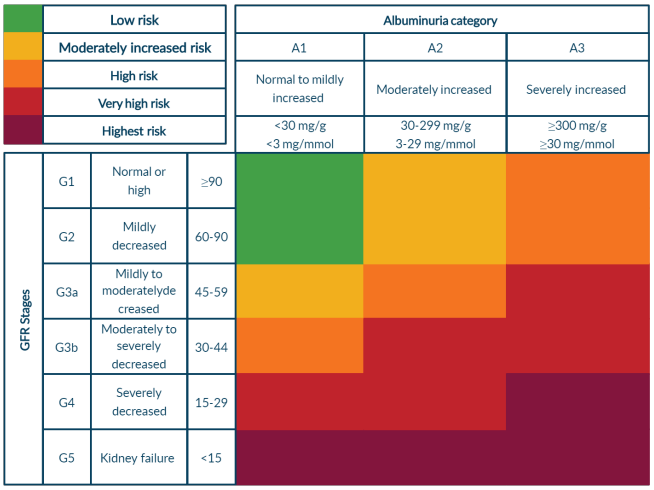

Fig. 1. Heat map of the severity.

To predict positive CKD status and the stages of CKD machine learning can be used. [6] Machine Learning grabs a major part of artificial intelligence when it comes to doing predictions from previous data using classification and regression methods. Application of machine learning methods to predict CKD has been explored based on multiple data sets. Among them, the dataset from UCI repository [7] (referred to as UCI dataset hereafter) is identified as a benchmark dataset [8]–[10]. Similar to most of the related work, this work considers the mentioned benchmark dataset.

When analysing clinical data related to CKD, if there are instances with missing attributes then the missing values handling method should be determined based on the randomness of the way they were missed. Moreover, the UCI data-set [7] has 400 instances which are a comparatively small number of samples with 25 attributes. In this case, the data set may have redundant (highly co-related) features or the data set does not represent all possibilities.

Thereby, this work identifies the limitations in handling missing values when analysing CKD data, proposes a new method to handle missing values and presents the evaluation of different methods based on UCI dataset. Further, this work also highlights the importance of statistical analysis as well

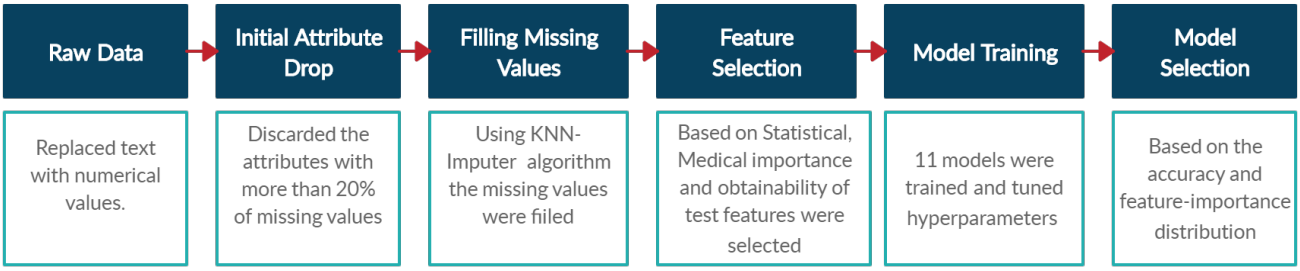

Fig. 2. Proposed workflow.

as the domain knowledge of the features when making a prediction based on clinical data related to CKD.

II. Related Work

A. J. Hussain and the team have obtained an accuracy of 0.995 in predicting CKD in early stages using multi-layer perception while including preprocessing of data set with neural networks to fill the missing values. The work-flow includes discarding the outliers, selecting the optimal seven attributes with statistical analysis, and discarding the attributes which have a higher inter co-relation by principal component analysis (PCA) [9]. In the mentioned work, the missing value filling algorithm has a significant impact on the accuracy of the trained models. However, because of using Neural Network for 20 attributes only with 260 fully completed data instances, the accuracy of missing value prediction is slightly reduced [9]. Discarding attributes which have more than 20% missing values has made a huge effect on accuracy in replacing missing values. The categorization of attributes according to the source such as blood test, urine test has assisted in selecting attributes for the training model from each category.

In terms of the five stages of CKD, a method has been proposed to predict a CKD stage with the highest accuracy for a stage with 0.997 and overall accuracy of 0.967 while discarding instances with missing values and by calculating the eGFR using the above data set with additional attributes about gender and race [11].

The slightly lower accuracy in the model causes the use of constants to replace missing values. However, in our work, it is shown that the randomness of missing data points is perfect according to Little's MCAR algorithm [12] (refer to Methodology section). Moreover, in [10], when consid-ering the features, the feature importance is biased towards serum creatinine. However in early stages of CKD serum creatinine can stay at regular values and all other features combined importance may not exceed the importance of serum creatinine [13], thereby making use of serum creatinine ineffective in disease prediction. The lack of inclusion of domain knowledge leads to question the reliability of the trained models when predicting the new instances outside the data set.

In 2017 a team of researchers has used 14 attributes to predict the CKD and achieved 0.991 accuracy with multiclass decision forest [8]. They have discarded the instances with missing values and trained a neural network and a logistic regression model which gave 0.975 and 0.960 of overall accu-racies respectively. The co-relations of the selected attributes vary in the range of [0.2 to 0.8]. Considering the medical perspective of attributes, hypertension can cause CKD as well as CKD causes hypertension and specific gravity has 0.73 co-relations to the class. Discarding such attributes could lower accuracy. . In 2015 Lambodar J. and Narendra Ku. K. have experimented with 8 machine learning models using WEKA data mining tool [14]. The highest Receiver Operating Characteristic (ROC) and accuracy were given by Naive Bayes, Multi-layer Perception and J48 algorithms as ROC of 1 and accuracies of 0.950, 0.9975 and 0.99 respectively. In the mentioned work, Kappa Statistics is used to find the argument strength and it has given the highest of 0.9947 for multilayer perceptron, 0.9786 as the next highest for decision table and J48 algorithms.

Considering the related work based on UCI CKD data set [7], it was observed that the reasons for many to have less accuracy are the poor handling of missing values and the method of attributes selection.

III. Methodology

The proposed methodology consists of 3 key steps: Data preprocessing, models training and model selection (Fig. 2).

A. Data Preprocessing : Missing Value Handling

In this work, data preprocessing was done in 2 steps. Firstly, the attributes having more than 20% data with missing values were filtered out(see Table I). Accordingly, the set of features, (red blood cells, sodium, potassium, white blood cell count, red blood cell count) is excluded in the analysis. The second step in data preprocessing handled the missing values in the remaining data.

In the pre-processing step, missing values have to be handled based on their distributions to achieve reasonable accuracy. In this work, to confirm the randomness of missing values, Little's MCAR test was performed. The potential bias due to missing data depends on the mechanism causing the data to be missing. The analytical methods applied to amend the missingness [15] are tested using the chi-square test of MCAR for multivariate quantitative data. It tests whether there exists a significant difference between the means of different missing-value patterns.

TABLE I
TESTS FOR MEASURING MULTIPLE ATTRIBUTES AND MISSING VALUE PERCENTAGE

Attribute	Missing percentage	Test to obtain
Class	0.00 %	
Appetite	0.25%	Doctor's Inspection
Pedal Edema	0.25 %	Doctor's Inspection
Anemia	0.25 %	FBC
Hypertension	0.50 %	Doctor's Inspection
Diabetes Mellitus	0.50%	FBC
Coronary Artery Disease	0.50 %	Doctor's Inspection
Pus Cell Clumps	1.00 %	UFR
Bacteria	1.00 %	UFR
Age	2.2 5 %	Doctor's Inspection
Blood pressure	3.00 %	Doctor's Inspection
Serum creatinine	4.25 %	SERUM CREATININE
Blood Urea	4.75 %	BLOOD UREA
Blood Glucose Random	11.00 %	RBS
Albumin	11.50 %	UFR
Specific Gravity	11.75 %	UFR
Sugar	12.25 %	UFR
Hemoglobin	13.00 %	FBC
Pus Cell	16.25 %	UFR
Packed Cell Volume	17.50 %	FBC
Sodium	*21.75 %*	*SERUM ELECTROIDES*
Potassium	*22.00 %*	*SERUM ELECTROIDES*
White Blood Cell Count	*26.25%*	*FBC*
Red Blood Cell Count	*32.50 %*	*FBC*
Red Blood Cells	*38.00 %*	*UFR*

Note: Text in *italics* corresponds to attributes excluded from the analysis and the tests to do were obtained from doctors

TABLE II
LITTLE'S MCAR TEST RESULTS

Name	Value
Chi.Square	3160.494
degree of freedom	2164
P value	0
Missing patterns	105

Based on Little's MCAR test results as shown in Table II, the missing values were considered to be completely random as the 'p' value equals zero. Accordingly, substituting missing values with a constant may drop the accuracy and bias the prediction process since there are more positive CKD instances than negative CKD instances. This scenario can be identified in some related work [8], [16].

By considering the the drawback (of [9]) as mentioned in the related work, the K Nearest Neighbor Imputer algorithm was used in this work to fill the missing values.

When applying the algorithm, the original distribution of the dataset has been maintained by using the number of estimators(a hyperparameter of the algorithm) as same the number of complete instances, which gave the minimum mean and minimum standard deviation change.

B. Data Preprocessing : Feature Selection

The absolute values of heat map of correlations of attributes to the class label (Fig. 3) show that hemoglobin, specific gravity, albumin, hypertension and diabetes mellitus have the highest correlations (more than 0.5). Then the secondary attributes pus cell, blood glucose random, appetite, blood urea, pedal edema, sugar, anemia and serum creatinine are the attributes which have correlations of more than 0.3.

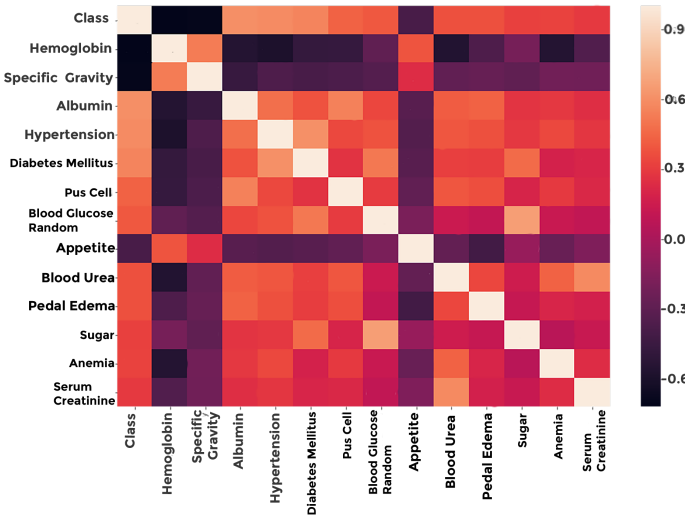

Fig. 3. Heat map of the co-relation of attributes to the class variable.

Thereafter, considering the distribution of attributes values and the medical perspective of the attributes specific gravity, albumin, haemoglobin, hypertension, diabetes mellitus, blood glucose random and serum creatinine were selected as the optimal subset of attributes to predict CKD. The selection of the mentioned attributes is explained in detail below.

Specific gravity and albumin has only 5 sets of values in each (Fig. 4). When plotted against each other, their values forms a distinct cluster with CKD negative instances.

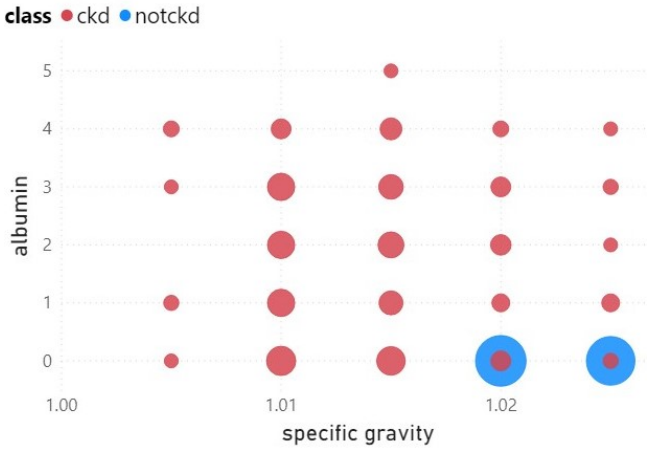

Fig. 4. Distribution of albumin over specific gravity.
Note: The count of each type represented by the size of the circle.

The amount of albumin is estimated based on a test for protein in the urine. An excess amount of protein in urine implies that the kidney's filtering units have been damaged by disease or due to fever or heavy exercise. Many tests should be done to confirm the condition over several weeks.

Generally, the hemoglobin level can decrease due to three reasons, namely decreased red blood cell production, increased red blood cell destruction and blood loss. Healthy kidneys produce a hormone called erythropoietin (EPO) [17].

TABLE III
PERCENTAGE CHANGE OF STATISTICS OF ATTRIBUTES AFTER FILLING MISSING VALUES

	Hb	Specific Gravity	Albumin	Hypertension	DM	Pus Cell	Blood Glucose Random	Appetite	BU	Pedal Edema	Sugar	Anemia	SC
count	13.00	11.75	11.50	0.50	0.50	16.25	11.00	0.25	4.75	0.25	12.25	0.25	4.25
mean	0.66	0.01	-1.39	-0.34	-0.36	-2.22	-0.14	0.04	-0.22	-0.10	-4.53	-0.15	-0.43
std	-6.32	-6.15	-5.95	-0.19	-0.19	-8.94	-5.86	-0.12	-2.43	-0.12	-6.30	-0.12	-2.16
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	5.29	0.49	0.00	0.00	0.00	0.00	1.98	0.00	0.00	0.00	0.00	0.00	0.00
50	2.62	-0.12	100	0.00	0.00	0.00	3.97	0.00	4.55	0.00	0.00	0.00	7.14
75	-2.56	0.00	0.00	0.00	0.00	100	-2.56	0.00	-3.00	0.00	100	0.00	0.89
max	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

A hormone is a chemical produced by the body and released into the blood to help trigger or regulate particular body functions. EPO prompts the bone marrow to make red blood cells, which then carry oxygen throughout the body. When kidneys are diseased or damaged, they do not make enough EPO. As a result, the bone marrow makes fewer red blood cells, causing anaemia but before it causes anaemia (which happens after less than 50% of one kidney is properly functions), the haemoglobin levels change slightly.Furthermore, the plot of hemoglobin vs serum creatinine also shows a separation of the two classes: positive and negative (Fig. 5) [2].

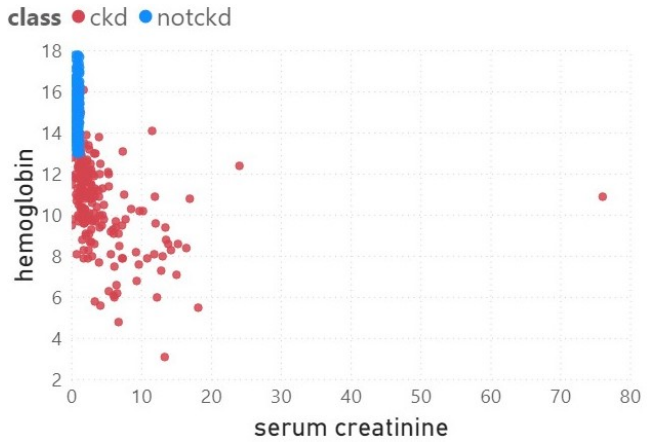

Fig. 5. Distribution of hemoglobin over serum creatinine.

Serum creatinine is also known as creat, blood creatinine or creatinine. Creatinine is a waste product produced by muscles from the breakdown of a compound called creatine. Creatinine is removed from the body by the kidneys. This test measures the amount of creatinine in the blood. creatine is part of the cycle that produces energy needed to contract muscles. Both creatine and creatinine are produced by the body at a relatively constant rate. Apart from issues directly related to kidney, a high-protein diet, congestive heart failure, complications of diabetes and dehydration can also increase the level of Creatinine in the blood. The normal range of Creatinine is 0.6 –1.1 mg/dL in women and 0.7 –1.3 mg/dL in men.

Two main causes of chronic kidney disease are diabetes and high blood pressure(see Fig. 6), which are responsible for up to two-thirds of the cases. Diabetes causes damage to many organs in the body, including the kidneys, heart, blood vessels, nerves and eyes. High blood pressure, or hypertension, occurs when the pressure of blood against the walls of blood vessels increases. If uncontrolled, or poorly controlled, high blood pressure can be a leading cause of heart attacks, strokes and CKD. Nevertheless, CKD can lead to high blood pressure.

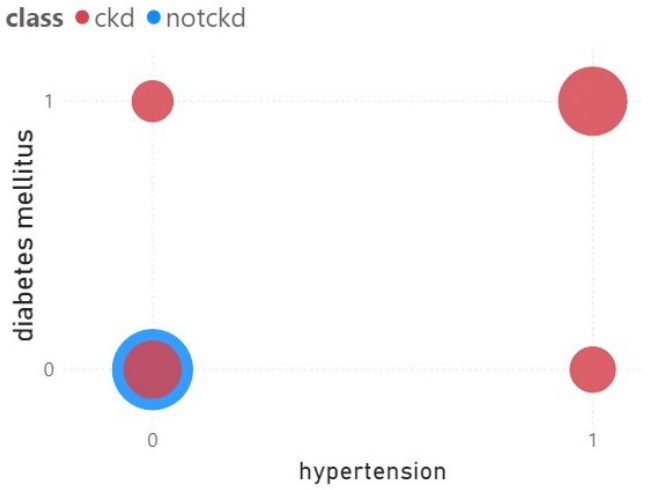

Fig. 6. Distribution of diabetes mellitus over hypertension.

In addition above mentioned factors, feasibility and the obtainability (Table I) was also considered in attributes selection.

The distribution of appetite values against the class shows that it biases towards good appetite (Fig. 7). However, CKD is not the only reason to have a poor appetite, which will mislead the predictions when applying the trained model to a new scenario.

C. Model Training

In this work, 11 classification models were considered in training. They are logistic regression, k-Nearest Neighbors (KNN) regression, SVC with a linear kernel, SVC with RBF kernel, Gaussian NB, decision tree classifier, random forest classifier, XGB classifier, extra trees classifier, an ada boost classifier and a classical neural network. The dataset was divided into 3 parts as 70% training data, 15% cross-validation data and 15% test data randomly. The models were further optimized by hyperparameter tuning from a genetic algorithm and grid search for the training dataset.

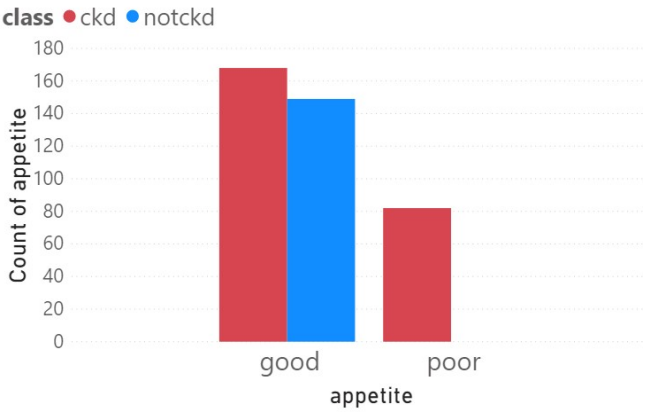

Fig. 7. The distribution of appetite.

From the mentioned 11 algorithms, 6 algorithms outperformed in training accuracy, testing accuracy and in cross-validation accuracy. Those are the decision tree classifier, random forest classifier, XGB classifier, extra trees classifier, ada boost classifier and classical neural network.

The implementations and evaluation were done using Python Sci-kit, and Keras frameworks.

TABLE IV
ACCURACIES OF EACH ALGORITHM

Algorithm	Training accuracy	Cross validation accuracy	Testing accuracy
Decision Tree Classifier	100.00%	100.00%	100.00%
Random Forest Classifier	100.00%	100.00%	100.00%
XGB Classifier	99.28%	100.00%	100.00%
Extra Trees Classifier	100.00%	100.00%	100.00%
Ada Boost Classifier	100.00%	100.00%	100.00%
KNN	97.85%	98.33%	98.33%
Classical Neural Network	97.81%	97.50%	97.50%
SVC Linear	97.14%	96.66%	96.66%
Logistic Regression	96.07%	96.66%	95.00%
SVC RBF	94.64%	95.00%	95.00%
Gaussian NB	95.35%	95.00%	93.33%

TABLE V
PRECISION, RECALL AND F1-SCORE OF EACH ALGORITHM

Algorithm	Precision	Recall	F1-Score
Decision Tree Classifier	1.000	1.000	1.000
Random Forest Classifier	1.000	1.000	1.000
XGB Classifier	1.000	1.000	1.000
Extra Trees Classifier	1.000	1.000	1.000
Ada Boost Classifier	1.000	1.000	1.000
KNN	1.000	0.975	0.987
Classical Neural Network	0.962	1.000	0.981
SVC Linear	1.000	0.950	0.974
Logistic Regression	1.000	0.925	0.961
SVC RBF	1.000	0.925	0.961
Gaussian NB	0.973	0.925	0.948

D. Model Evaluation and Selection

Based on the results (Table IV, V), the algorithms which resulted in the highest accuracy in all 3 data sets were selected. Those are decision tree classifier, random forest classifier, XGB classifier, extra trees classifier and ada boost classifier.

Even though the mentioned models gave 100% accuracy, its important to identify attributes with the highest impact on each model to make the decision (Table VI).

After identifying the importance of selected features for each prediction algorithm, the standard deviation of the feature importance of each algorithm was calculated, which explicitly showed the algorithm's bias towards different attributes (Table VI, Table VII, Fig. 8).

The results (Table VI, Table VII, Fig. 8) show that the extra trees classifier has the lowest bias towards features next to the random forest classifier. decision tree classifier has the highest bias out of all.

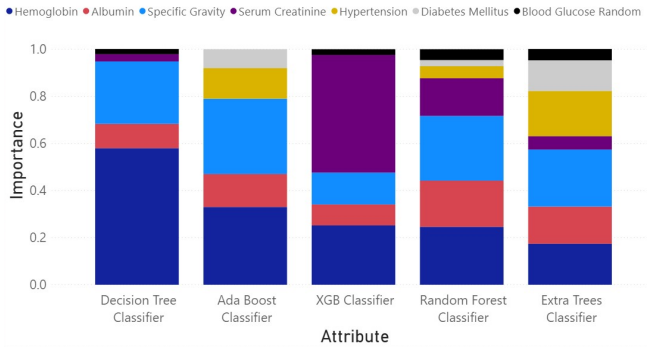

Fig. 8. Feature importance of each trained model.

IV. DISCUSSION

The data distribution has properly covered the whole domain in CKD, but the general attributes like appetite, anaemia and pedal oedema are biased towards CKD. It is easy to achieve an accurate prediction using this data set but in the general context, it may lead to false positives as observed in the recall column of Table V. Further, the missing values which were completely missed at random made it impossible to achieve a perfect accuracy without filling them from a collaborative imputer instead of a constant. Considering the medical importance of the attributes, some of them have a lesser co-relation compared to others because of the stage they appear in the patient. When training the models it makes a huge impact on the accuracy. After training the model, it clearly shows that tree structures have higher accuracy than other classification algorithms, which can be justified from the distribution of the data set since the selected attributes have a clearer separation in the class except for serum creatinine attribute. Finally, when selecting the algorithm, some trained models have a bias towards some attributes as shown in the Table VI, Considering the causes of change of the nominal values of them, it has many different possibilities apart from CKD. Therefore, it motivates to rely less on one attribute and consider more when making the decision and based on that the extra tree classifier has been selected.

V. CONCLUSION AND FUTURE WORK

Chronic Kidney Disease (CKD) is a life-threatening issue that affects almost 14% of the world population and predicting it with a 100% overall accuracy makes it possible for

TABLE VI
FEATURE IMPORTANCE OF EACH ALGORITHM

Attribute	Decision Tree Classifier	Random Forest Classifier	XGB Classifier	Extra Trees Classifier	Ada Boost Classifier
Hemoglobin	0.580	0.246	0.252	0.174	0.330
Specific Gravity	0.265	0.275	0.135	0.242	0.320
Serum Creatinine	0.031	0.160	0.500	0.057	0.000
Albumin	0.103	0.196	0.089	0.158	0.140
Hypertension	0.000	0.051	0.000	0.192	0.130
Diabetes Mellitus	0.000	0.026	0.000	0.130	0.080
Blood Glucose Random	0.022	0.046	0.024	0.048	0.000

TABLE VII
STANDARD DEVIATION OF FEATURE IMPORTANCE OF ALGORITHMS

Attribute	Extra Trees Classifier	Random Forest Classifier	Ada Boost Classifier	XGB Classifier	Decision Tree Classifier
Standard Deviation	0.070684746	0.102219419	0.136224604	0.181369263	0.214295746

people to get to know it in the early stages to get treated with a minimum cost and minimum risk. Proper feature engineering helps to reduce the number of features needed for the prediction algorithm and practically it reduces the number of medical tests to be taken. Filling missing values based on the distribution of them along with the collocation of other attributes by K Nearest Neighbors-imputer (KNN-imputer) instead of replacing with a constant directly leads to higher accuracy in prediction models compared to the related work done with the same dataset. Furthermore, the extra trees classifier and the random forest classifier are the better algorithms to do the predictions for CKD since those have 100% overall accuracy and minimal bias towards specific attributes compared to other models. This work suggests a new workflow including data preprocessing, missing values handling and features selection to predict CKD status as positive or negative. Furthermore, this work highlights the importance of incorporating the domain knowledge into feature selection when analysing clinical data related to CKD.

Accordingly, it is worthwhile to explore the use of KNN-imputer based approach to handle missing values in data sets related to multiple diseases in future. Furthermore, more insights into CKD can be gained by adding knowledge of genomics, water consumption patterns and food types into the analysis.

ACKNOWLEDGEMENT

We thank Ms. R. I. Vithanage (Faculty of Engineering, University of Peradeniya) for the comments to improve the manuscript.

REFERENCES

[1] "Your Kidneys & How They Work | NIDDK." [Online]. Available: https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work

[2] "Kidney Disease: The Basics," Aug. 2014. [Online]. Available: https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics

[3] "Global Facts: About Kidney Disease," [Online]. Available: https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease/. [Accessed: 20-Feb-2020].

[4] "Estimated Glomerular Filtration Rate (eGFR)," Dec. 2015. [Online]. Available: https://www.kidney.org/atoz/content/gfr

[5] F. E. Murtagh, J. Addington-Hall, P. Edmonds, P. Donohoe, I. Carey, K. Jenkins, and I. J. Higginson, "Symptoms in the month before death for stage 5 chronic kidney disease patients managed without dialysis," *Journal of pain and symptom management*, vol. 40, no. 3, pp. 342–352, 2010.

[6] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *Journal of translational medicine*, vol. 17, no. 1, p. 119, 2019.

[7] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[8] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2017, pp. 291–296.

[9] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2018, pp. 1–9.

[10] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019.

[11] C. A. Johnson, A. S. Levey, J. Coresh, A. Levin, and J. G. L. Eknoyan, "Clinical practice guidelines for chronic kidney disease in adults: part 1. definition, disease stages, evaluation, treatment, and risk factors," *American family physician*, vol. 70, no. 5, pp. 869–876, 2004.

[12] C. Li, "Little's test of missing completely at random," *The Stata Journal*, vol. 13, no. 4, pp. 795–809, 2013.

[13] S. Nair, S. V. O'Brien, K. Hayden, B. Pandya, P. J. G. Lisboa, K. J. Hardy, and J. P. H. Wilding, "Effect of a Cooked Meat Meal on Serum Creatinine and Estimated Glomerular Filtration Rate in Diabetes-Related Kidney Disease," *Diabetes Care*, vol. 37, no. 2, pp. 483–487, Feb. 2014. [Online]. Available: http://care.diabetesjournals.org/cgi/doi/10.2337/dc13-1770

[14] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 02, pp. 193–198, 2017.

[15] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts," *BMC medical research methodology*, vol. 17, no. 1, p. 162, 2017.

[16] S. Vijayarani, S. Dhayanand *et al.*, "Data mining classification algorithms for kidney disease prediction," *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 4, pp. 13–25, 2015.

[17] "Facts About Chronic Kidney Disease," May 2020. [Online]. Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease