Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics

Ahmed J. Aljaaf[1, 2], Dhiya Al-Jumeily[2], Hussein M. Haglan[1], Mohamed Alloghani[3], Thar Baker[2], Abir J. Hussain[2], and Jamila Mustafina[4]

[1]Computer Centre, University of Anbar, Ramadi-Anbar, Iraq
[2]Applied Computing Research Group, Faculty of Engineering and Technology, LJMU, Liverpool, L3 3AF, UK
[3]Abu Dhabi Health Services Company (SEHA), Abu Dhabi, UAE
[4]Kazan Federal University, Kazan, Russia
{A.J.Kaky, D.Aljumeily, A.Hussain}@ljmu.ac.uk; M.AlLawghani@2014.ljmu.ac.uk;
Hussein.m.haglan@uoanbar.edu.iq; DNMustafina@kpfu.ru;

***Abstract*—Chronic Kidney Disease is a serious lifelong condition that induced by either kidney pathology or reduced kidney functions. Early prediction and proper treatments can possibly stop, or slow the progression of this chronic disease to end-stage, where dialysis or kidney transplantation is the only way to save patient's life. In this study, we examine the ability of several machine-learning methods for early prediction of Chronic Kidney Disease. This matter has been studied widely; however, we are supporting our methodology by the use of predictive analytics, in which we examine the relationship in between data parameters as well as with the target class attribute. Predictive analytics enables us to introduce the optimal subset of parameters to feed machine learning to build a set of predictive models. This study starts with 24 parameters in addition to the class attribute, and ends up by 30% of them as ideal sub set to predict Chronic Kidney Disease. A total of 4 machine learning based classifiers have been evaluated within a supervised learning setting, achieving highest performance outcomes of AUC 0.995, sensitivity 0.9897, and specificity 1. The experimental procedure concludes that advances in machine learning, with assist of predictive analytics, represent a promising setting by which to recognize intelligent solutions, which in turn prove the ability of predication in the kidney disease domain and beyond.**

***Index Terms*—Chronic Kidney Disease; Predictive analytics; Machine learning;**

I. Introduction

Chronic Kidney Disease (CKD) is a serious public health condition worldwide that tied to unpleasant health outcomes, particularly in low-to-middle income countries where millions die due to lack of affordable treatment [1] [2]. CKD is a long-term condition induced by damage to both kidneys. Kidney damage refers to any kind of kidney pathology that gives the possibility to reduce the capacity of kidney functions, particularly the reduction in glomerular filtration rate (GFR) [4]. Kidneys have millions of tiny blood vessels work as filters to remove waste products from the blood. In some instances, this filtration system breaks down and kidneys lose their ability to filter out waste products, which results in kidney disease. There is no single underlying cause to CKD, but the deterioration is commonly irreversible and can lead to serious health problems. In the past decade, the US National Kidney Foundation's Kidney Disease Outcomes Quality Initiative has established the first guideline that defined CKD, regardless of the cause, as based on 3 or more months of either kidney damage (pathologic abnormalities, or imaging abnormalities) or GFR rate <60 mL/min/1.73 m2 [3].

The mean prevalence rate of CKD in both of US and Europe is about 11% [4]. The economic burden of CKD to health care systems is enormous, and tends to increase because of aging populations and elevated prevalence of Type 2 diabetes [5]. Between 2011 and 2012, there were about 1.9 million adults with CKD in England as registered in the Quality Outcomes Framework (QOF). However, the prevalence of category G3-G5 CKD in England has been anticipated to 6.8%, which equates to 3 million individuals [6]. The overall annual cost of CKD to the UK National Health Service (NHS) is estimated at about £795 for every patient diagnosed with CKD and recorded in the QOF; this is equivalent to 1.44 to 1.45 billion a year [5]. Early detection and treatment of CKD can possibly slow or stop the progression of kidney disease [1] [4]. Identifying people with CKD in early stage will help reduce the risk of end-stage renal disease (ESRD) [5].

Epidemiology discloses relationships between the development of CKD and many other clinical characteristics. Factors that can influence the development of CKD consist of genetics, diabetes, hypertension, and ageing [4]. In general, a nephrologist uses two tests to check for CKD, blood test and urine test. The blood test measures how well kidneys are filtering the blood to remove creatinine, which is a normal waste product of muscle breakdown. The urine test, on the other hand, can show the existence of protein in the urine. Protein in particular (albumin) is a component of the blood that does not normally pass through the kidney filters into the

978-1-5090-6017-7/18/$31.00 ©2018 IEEE

urine. If urine test reveals the existence of albumin, it means that the kidney filters are damaged and may reflect Chronic Kidney Disease.

This study uses CKD data set donated by Apollo Hospitals, India that available online at the UCI machine learning repository [5]. Using this data, we aim at (a) examine the correlation between predictors (i.e. input parameters) and the development of CKD using predictive analysis approaches. This will enable us to reduce the number of required parameters to predict the occurrence of CKD as well as eliminate redundant and noisy parameters. (b) Examine the capability of using one of the two tests for the prediction of CKD, either blood test or urine test, and then measure its accuracy and applicability. (c) Employ machine learning methods to early prediction of CKD using the most relevant and representative parameters.

This paper is organised as follows. Section II presents the data and its processing methods, which starts from handling extreme values to the use of predictive analytics for selecting optimal subset of parameters. Section III builds four predictive models for CKD and examine them using a number of performance matrices. Section IV discusses key differeances between models and concludes the study. Limitations is given in Section V.

II. Materials and Methods

A. Data

The CKD data set consists of 24 parameters (i.e. predictors) in addition to the binary class attribute. As illustrated in table 1, parameters are distributed as three main groups. Parameters extracted from blood serum chemistry and blood haematology tests, which is about 41.7%. Parameters derived from urine test represent about 29.15%. The last group of parameters includes general information about other clinical factors that may induce CKD and represents 29.15%. Total number of records in this data set is 400, in which 62.5% are for patients diagnosed with CKD, while other 37.5% are for healthy individuals. There are 12 numerical parameters, two categorical with five levels, while the remaining parameters are binary and been coded as zero for normal instances and one for abnormality. CKD data set is a raw data and we therefore consider a number of data processing techniques before prior to analysis and the development of predictive models.

TABLE I: DATA DESCRIPTION

Parameters	Measurement	Missing	Percent
Glucose	Num. (mg/dL)	44	11
Urea	Num. (mg/dL)	19	4.8
Creatinine	Num. (mg/dL)	17	4.3
Sodium	Num. (mEq/L)	87	21.8
Potassium	Num. (mmol/L)	88	22
Haemoglobin	Num. (g/dL)	52	13
Packed Cell Volume	Num.	71	17.8
White Blood Cell Count	Num. (cells/mcL)	106	26.5
Red Blood Cell Count	Num. (m.c./mcL)	131	32.8
Specific Gravity	Num. (1.002-1.030)	47	11.8
Urine Glucose	Category (0-5)	49	12.3
Albumin	Category (0-5)	46	11.5
Bacteria	Binary	4	1
Red Blood Cells in Urine	Binary	152	38
Pus Cell	Binary	65	16.25
Age	Num. (years)	9	2.3
Hypertension	Binary	2	0.5
Blood Pressure	Num. (mm/Hg)	12	3
Diabetes	Binary	2	0.5
Coronary Artery Disease	Binary	2	0.5
Appetite	Binary	1	0.25
Pedal edema	Binary	1	0.25
Anemia	Binary	1	0.25

B. Extreme values

Extreme values or outliers are extreme data points that located away from other members of a given data cluster [8]. In CKD data set, outliers may be arise because of errors or a natural variance of data. These extreme data points usually increase the variance of data and influence the normal distribution assumption that is required for parametric analysis [9]. Researchers usually use boxplots as an easy way of visual inspection to detect outliers; however, we have employed the following mathematical notation using interquartile range (IQR) to identify extreme data points in CKD.

$$Ext.\ values = \begin{cases} Points > Q3 + 1.5\,(IQR) \\ Points < \ Q1 - 1.5\,(IQR) \end{cases} \quad (1)$$

Where $Q1$ is the first quartile, $Q3$ is the third quartile, and $IQR = Q3 - Q1$ [9]. We have three main option to handle outliers, (a) keep and handle them just like any other data points. (b) Remove them from data sample, and (c) modify them to the next highest or lowest values within the distribution that are not suspected to be outlier. Modifying outliers is the recommended method [9], in particular with CKD data set as the data points are most likely to be legitimate data points. Therefore, we have decided to modify extreme outliers only, which represent 0.24% of the total data points. We believe that modifying this tiny amount of extreme outliers would not have a noticeable impact on the statistical inference, while it would make these extreme data points closer to the population sample. Ghosh and his colleague in [10] have reported that modifying up to 2.5% of outliers would probably maintain characteristics of the data and would not adjust the distribution considerably. Furthermore, this process has dropped the skewness of the following parameters, blood pressure by 66%, blood glucose by 12%, blood urea by 24%, and serum creatinine by 65%.

C. Incomplete cases

Incomplete cases or missing values is one of the popular problems in real-world data sets, and especially in medical data [21]. Approximately 45% of all data sets in the UCI online machine-learning repository have some forms of missing as

reported by Tran and his partners [11]. Table 1 reveals the mount of incomplete cases to CKD data parameters. We can observe that incomplete cases vary from a parameter to another. It starts as little as 1%, and reaches 38% for red blood cells (rbc) parameter. Incomplete cases can cause serious concerns for the development of predictive models, including the non-applicability of several machine-learning (ML) methods to data with incomplete cases [9]. Even though some ML methods can handle this type of data by ignoring them, however the majority cannot. Thus, waste of data and fundamental learning errors are presumably take place. Therefore, the first step toward valid predictive models is to address incomplete cases.

So let us represent CKD data set as a matrix $N \times P$ that contain data values of P parameters for all N participants. A parameter $P = (P_1, P_2, ..., P_j)$, where j is the dimension of data set (i.e. 400). The complete cases (i.e. observed values) in certain parameter P_i are collectively denoted as P_i^{obs}, while incomplete cases (i.e. missing values) of P_i are collectively denoted as P_i^{miss}. Hence $P = (P_i^{obs}, P_i^{miss})$. In this study, we use multiple imputations (MI) method [12] to handle incomplete cases of parameters that meet our threshold of missingness according to the following equation [9].

$$\forall\ P_i \in P = \begin{cases} Impute, & P_i^{miss} < R \\ Ignore, & otherwise \end{cases} \tag{2}$$

In this context, we will discard any data parameter P_i that has incomplete rate of greater than or equal to R, where $R = (1/5)\,N$ or 20% of the whole population. Consequently, five parameters have been ignored and will not been involved in the development of predictive models. Theses parameters are sodium, potassium, red and white blood cells count, and finally red blood cells in urine. The method of ignoring parameters with incomplete rate of greater than or equal to R is proposed and discussed in our previous work [9]. In general, MI uses regression analysis to fill a model for incomplete cases on a multivariate basis, where MI treats parameters with incomplete cases as outcomes and the rest of parameters as predictors. MI is a sophisticated approach that perceives the uncertainty related to imputation process.

In MI method, the imputation process repeats m times, in this study we have specified $m = 5$. This generates five complete data sets and variations between imputed sets represent uncertainty in the imputation process. Finally, we analysis imputed sets separately to generate multiple analysis results (i.e. estimates such as mean, slandered deviation, and regression coefficients) using following equations [12].

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}^{(i)} \tag{3}$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \tag{4}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m} \left(\hat{Q}^{(i)} - \bar{Q}\right)^2 \tag{5}$$

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U^{(i)} \tag{6}$$

For m imputed sets, the estimate Q and the estimated total variance T are calculated as described in equation 3 and equation 4 respectively [13]. Where $\bar{Q}$ is the final combination of estimate Q, and $Q = (Q_1, \ldots, Q_k)$, which is the factor to be estimated with k elementns.In addition, $\hat{Q}^{(i)} = \left(\hat{Q}_1^{(i)}, \ldots, \hat{Q}_k^{(i)}\right)$ is the estimated factor using i^{th} set of imputed data, where $i = (1, \ldots, m)$. B and U are respectively the between-imputation and the average within-imputation of CKD data. Finally, $U^{(i)}$ is the estimated covariance matrix of $\hat{Q}^{(i)}$ [13].

D. Analysis

In this section, we employ predictive analytics methods to examine relevance of input parameters to class attribute, as well as association between parameters themselves. This is a fundamental point toward an effective and valid prediction of CKD. Typically, the stronger the relevance of a parameter to the class attribute means that this parameter is necessary for an optimal learning performance and predication. Conversely, parameters with weak relevance may not be important for the learning procedure, and we can discard them as noisy parameters. However, strong association between two parameters indicates the existence of redundant data that can be eliminated to reduce the number of input parameters. Therefore, it is worthwhile to analyse input parameters to define their discriminatory power in the prediction of CKD in the early stage. This step enables us to understand the level of overlap between CKD and healthy individuals with respect to certain parameter.

1) Parameters derived from blood tests: After declaring personal and family history, the first step toward the diagnosis of CKD doctor's take is usually to order blood serum chemistry test to measure kidneys function through the level of waste products such as creatinine and urea in the blood. There are many other parameters within blood test that can be used as indicators of high risk of CKD such as level of blood glucose, haemoglobin, and haematocrit (i.e. packed cell volume (PCV). In this section, we measure the correlation between these parameters using different statistical test such as Pearson's correlation, Chi-Square test for association, or analysis of variance (ANOVA). The main aim is to identify any correlation and its significant values to understand the relationships between these parameters as well as to verify the likelihood of redundant parameters.

Pearson's correlation test has revealed a strong positive relationship between creatinine and urea. The correlation coefficient of these two waste products was 0.801 at a significant level of 0.01. Creatinine is a waste product of muscle catabolism of creatine phosphate [14]. Although creatinine level can be raised by many factors such as age, sex, ethnicity, diet rich in proteins and muscle mass, which in turn influences GFR itself as a biomarker [4]. However, it remains the most

prescribed analyses to estimate the GFR [14]. On the other hand, urea is the main nitrogenous waste product of protein and amino acid catabolism. The use of urea as indicator of kidney function might not be accurate enough to influence GFR. In fact, GFR has to decline by around a half before urea level increases above the upper limit of the reference range as stated in medical literature [15]. Therefore, urea seems to be insensitive pointer of reduced GFR, and therefore cannot be used for an early predication of CKD. As shown in figure 1, we have scattered urea values by creatinine levels and can clearly observe the correlation between them. Values of these two parameters for healthy individuals are squeezed in the lower left-hand side, where values between 5 and 50 mg/dL for urea and mostly less than 3 mg/dL for creatinine. Consequently, we discard the use of urea as a predictor because it is a redundant parameter according to the correlation test as well as it is not sufficient for early prediction of CKD.

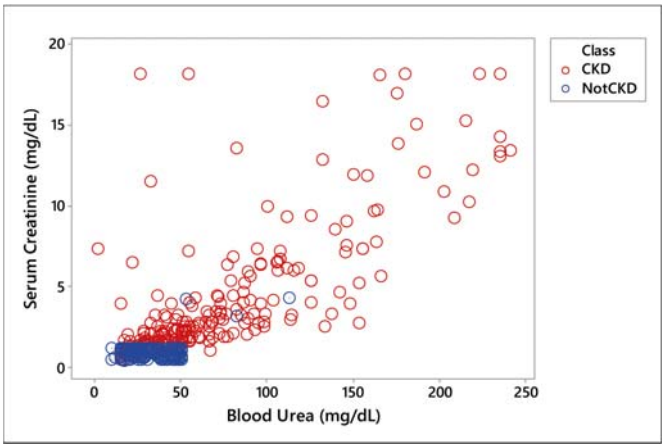

Fig. 1: Creatinine and urea in blood by class attribute.

Similarly, we have notes that the correlation between haemoglobin and PCV is 0.883 at a significant level of 0.01. This is a strong positive correlation as presented in figure 2, where values of these two parameters for healthy individuals aggregated at the top right-hand side of the scatter plot. Haemoglobin is a complex protein located in red blood cells, which contains an iron molecule. The primary role of haemoglobin is to deliver oxygen from the lungs to the body cells, and to substitute the oxygen for carbon dioxide. According to the Scottish Intercollegiate Guidelines Network (SIGN) [4], the kidney creates less erythropoietin gradually with the progression of CKD and patients can become anaemic. In addition, estimates reveal that one-third of males and two-third of females in stage 4 of CKD have a haemoglobin level below the normal range.

PCV is the volume percentage of red blood cells in blood, which increases with a rise in the quantity of red blood cells or a decline in the plasma volume. Haemoglobin and PCV might indicate that a patient has anaemia (figure 3), which is one of the indicator toward kidney failure. As can be observe in figure 2, the majority of patients with CKD (86%) recorded values of haemoglobin less than 13 g/dL,

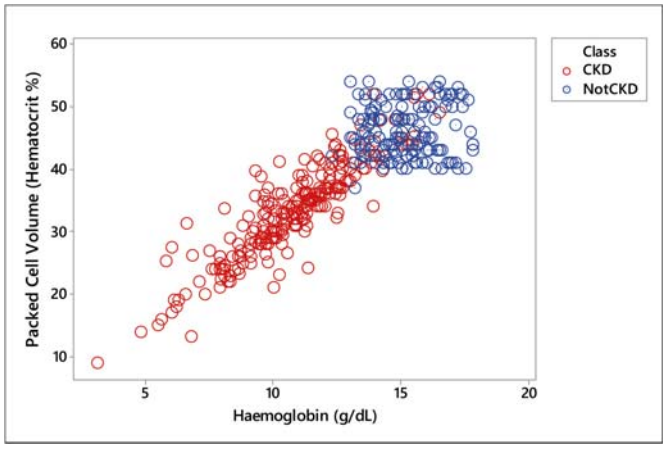

Fig. 2: PCV and haemoglobin in blood by class attribute.

compared to 0.66% of healthy individuals. Likewise only 6% of healthy individuals registered PCV level of below 40%, compared to 82.8% of patients with CKD. Moreover, ANOVA test with 95% of confidence interval has revealed a significant correlation between anaemia and both of haemoglobin and PCV at $p < 0.001$. Figure 3 illustrates the clear influence of anaemia on levels of PCV and haemoglobin. Based on these findings we can consider that the existence of PCV and anaemia parameters are redundant, which can be dispensed and thus dependence on the haemoglobin factor for early prediction of CKD.

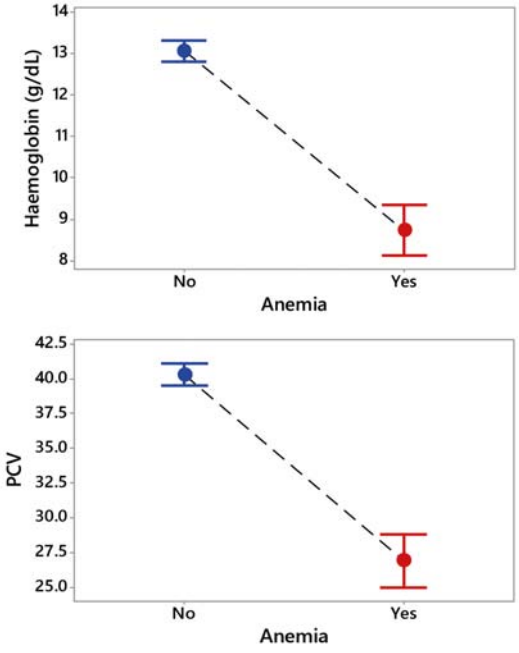

Fig. 3: The influence of anaemia on PCV and haemoglobin.

The final parameter in the blood test of CKD data is glucose; there is a small overlapping area between healthy individuals and patients with CKD as shown in below boxplot.

Although 2% of healthy individuals reported blood glucose of 140, compared to 57.2% of patients with CKD that reported blood glucose of 140 and higher. However, the overlapping area affects the correlation of blood glucose with the class attribute, which recorded at 0.427. Conversely, blood glucose appear to be strongly correlated to urine glucose with about 0.7 of correlation coefficient at significant level <0.001. Glucose levels in both of blood and urine are indicators of diabetes (0.520 at significant level <0.001). Therefore, it is not wise to use all three parameters together for predictive models, where using one of them is sufficient while others are redundant parameters.

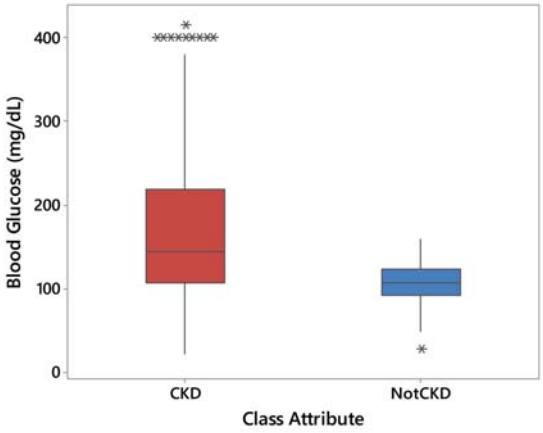

Fig. 4: Blood glucose of patients and healthy individuals.

2) Parameters derived from urine tests: Urine test can expose abnormalities that indicate to, and help to identify reasons of CKD. A 24-hour urine test reveals the amount of urine produced by kidneys. This provides a precise evaluation of how well the kidneys are functioning, as well as the amount of protein leaks from kidney's filter into the urine in a single day [16]. Six parameters are included within urine test section, namely specific gravity, urine glucose, albumin, existence of bacteria, pus cell, and pus cell clumps. Approximately, 80% of patients with CKD have recorded specific gravity of 1.015 and lower, whereas less than 1% of healthy individuals recorded same values. Specific gravity is one of the parameters that significantly correlated to class attributes, -0.720 at significant level of < 0.001.

We have discussed urine glucose parameters and its association to blood glucose and diabetes in the previous sub section. Moving on to albumin parameter, measuring albumin level in urea is essential for the diagnosis of kidney diseases, since it is proven risk factor for mortality of end-stage kidney disease in individuals having diabetes [17]. In CKD data, 99.3% of healthy individuals have a normal level of albumin, compared to 25.6% of patients with CKD at normal level. Moreover, albumin has a significant correlation to class attribute, 0.603 at a significant level of <0.001. Conversely, pus cells does not show a strong influence on class attribute. Although it is a sign

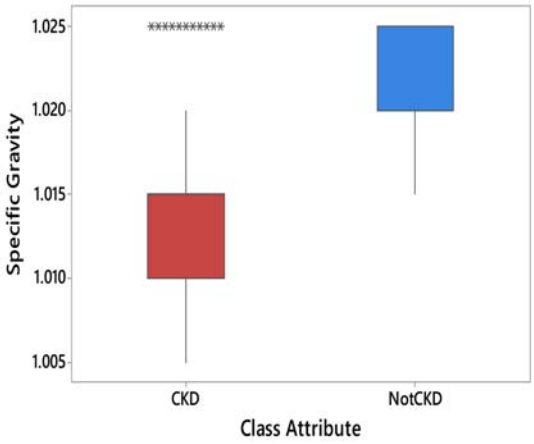

Fig. 5: Urine specific gravity.

of bacterial infection within the kidneys or bladder, however with correlation coefficient of 0.322, pus cells parameter may add less information to the end model. Similarly pus cell clumps parameter does not have a noticeable effect on class attribute with correlation coefficient of 0.265. It is not presented in about 83% of patients with CKD and 97% of healthy individuals. Pus cells and pus cell clumps parameters are both indicators of bacterial infection, where the existence of bacteria in urine as a parameter also does not show a significant contribution toward predicting class attribute. A correlation coefficient of bacteria parameter to class attribute is 0.187.

3) Other influencing parameters: Starting with age parameter, which is one of the factors that influence CKD. According to SIGN [4], CKD is observed more often in elder individuals and therefore is likely to increase the affected population as a whole. As shown in figure 6, the risk of developing CKD starts between 30 and 40 years of age, then sharply increases to reach its peak at about 60 years. ANOVA has revealed that there is a difference in the means age of population between patients with CKD and healthy individuals at significant level of <0.001. However, with a correlation coefficient of 0.215, age parameter does not appear to have a strong influence on class attribute.

High blood pressure is quite popular in CKD and represents a primary focus for intervention to prevent progression [4]. Jafar and other in [18], have reported that a systolic blood pressure of >130 $mmHg$ is significantly related to the development of CKD. In our data set, 34% of patients with CKD have reported a blood pressure of between 90 and 120 $mmHg$, while none of healthy individuals have recorded similar values. The correlation between blood pressure and class attribute is less than 0.3, which makes this parameter a weak for prediction of early stage of CKD. This is mainly because comparable quantities of patients have recorded a blood pressure of same levels. For example, about 29% of patients with CKD and 25.5% of healthy individuals have

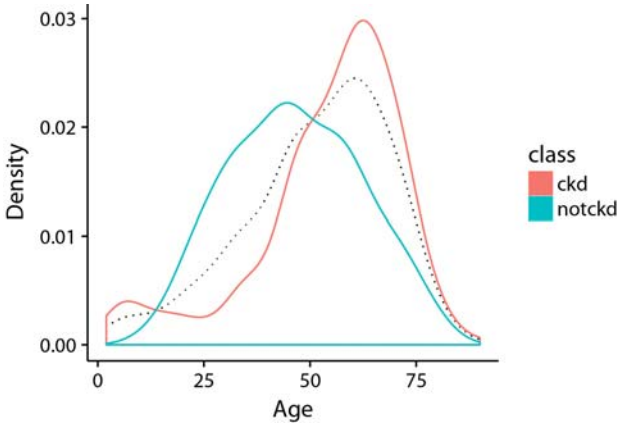

Fig. 6: Age distribution.

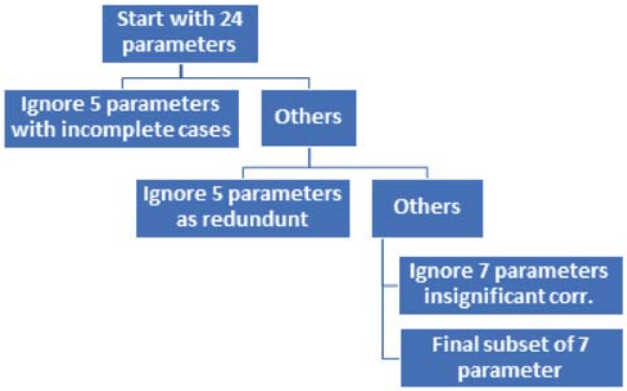

Fig. 7: Parameters selection procedure.

registered a blood pressure at 70 $mmHg$. Blood pressure can be an indicator of hypertension that has been collected as a binary parameter, i.e. presence or absence. However, with a correlation coefficient of 0.270, there is no significant influence between them. Conversely, we noticed a good correlation between hypertension parameter and class attribute, where none of healthy individuals have hypertension, compared to 41% of patients with CKD.

The prevalence of coronary artery disease (CAD) between patients with CKD is 13.6%, which is not a considerable to some extent to be considered for early prediction of CKD. Likewise for appetite parameter and pedal edema. The presence of coronary artery disease, appetite, and pedal edema was limited to a very small amount of patients, which reduces their ability to influence class attribute.

E. Optimal subset of parameters

In this stage, we get to choose the final sub set of parameters to create predictive models using a set of machine learning methods. As explained in figure 7, we started this study with 24 parameters and end up with 7 most informative and representative sub set of parameters. We have selected these 7 parameters by investigating their correlation to each other as well as to class attribute using different statistical methods. We have also identified threshold of correlation as a minimum boundary of parameter selection, by which we eliminate parameters with correlation coefficient of less than 0.300. Accordingly, the optimal subset of parameters includes haemoglobin, specific gravity, albumin, hypertension, blood glucose, creatinine, and pus cells. Moreover, we have employed a wrapper approach to confirm the selection of the final sub set of parameters. A majority vote of three machine-learning classifiers, i.e. random forest, support vector machine and logistic regression, has approved this selection as an optimal, representative sub set for predication of CKD.

III. Predictive models for CKD

As a final step before initialising predictive models, we normalised data to get all the seven parameters on the same level of measurement. Normalisation prevents parameters to overwhelm each other and enhance machine learning ability to measure similarities and distances between instances, and thus discover patterns in data. Moreover, Jin and others [19] have reported that normalised data are remarkably increasing the training speed of neural network. In this paper, we have employed a min-max normalisation method, which rescale parameters to be between 0 and 1 as formulated in the following equation [20].

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{7}$$

where x is a certain value to be normalised, x_{min} and x_{max} are the minimum and maximum observed values of a given quantitative attribute P_i, x_n is the normalised value of x, and $x,\ x_{min},\ x_{max} \in P_i$.

Now we reached the stage of assessing the capability of several machine learning (ML) methods for early prediction of CKD. The involved ML methods are classification and regression tree (RPART), support vector machine (SVM), logistic regression (LOGR) and multilayer perceptron neural network (MLP). We compare the overall performance of these ML methods using a number of performance matrixes including sensitivity (TPR), specificity (TNR), precision (PPV), and classification accuracy (ACC). The evaluation includes ROC curve analysis and measuring area under the curve (AUC), required time to build predictive models as well as overall error rate (ERR). Using holdout method, we have randomly divided the data set into 60/40 percent for training and testing.

The evaluation results of the predictive models for early prediction of CKD presented as follows. Table 2, lists experimental results of each model in terms of the six performance matrices, in addition to the overall error and required training time. Figure 8, provides a visual assessment of sensitivity, specificity and classification accuracy of ML methods. Figure 9, demonstrates the AUC values resulting from ROC analysis, PPV along with F1 measure as a harmonic mean of precision and sensitivity.

It is apparent that all of the predictive models have yielded considerably good results in predicting CKD, in which the

highest sensitivity was 0.9897 and has jointly achieved by MLP and LOGR models, followed by SVM and then RPART model. LOGR and MLP models have shown similar performance over nearly all of the performance matrices, with exception of the AUC, where MLP achieving the highest AUC values of 0.995. Although RPART model has obtained the highest specificity, however it was also the less sensitive to predict CKD. MLP and LOGR models were more stable with respect to performance analysis then RPART and SVM, and they are both showing lowest overall error rates.

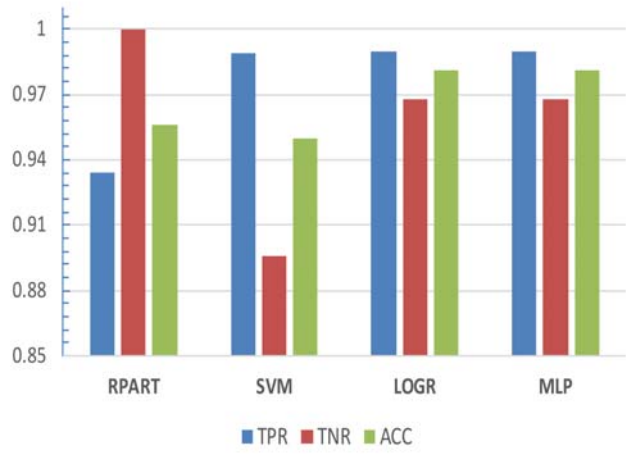

Fig. 8: Sensitivity, specificity, and accuracy.

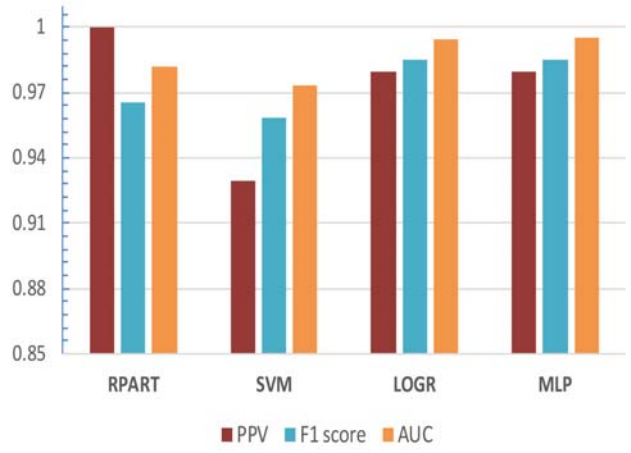

Fig. 9: Precision, F1, and Area under the ROC curve

According to the F1 score, as a harmonic mean of sensitivity and precision, RPART overcomes SVM that came at the end of the list, while the highest F1 score was 0.984 and shared between MLP and LOGR models. The ROC plot, i.e. TPR against 1-TNR, (figure 10) demonstrates the similarity in the performance profile with a few exceptions, while precision/recall curve (figure 11) shows some differences in predictive models behaviour. On this curve, SVM was clearly the least performed model, followed by LOGR model. The use of precision/recall curve is a common way to assess predictive

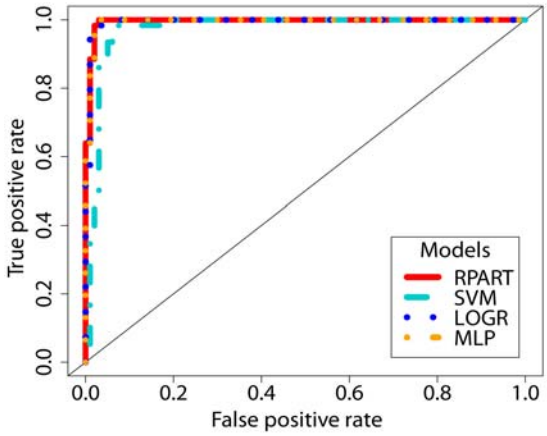

Fig. 10: Sensitivity, specificity, and accuracy.

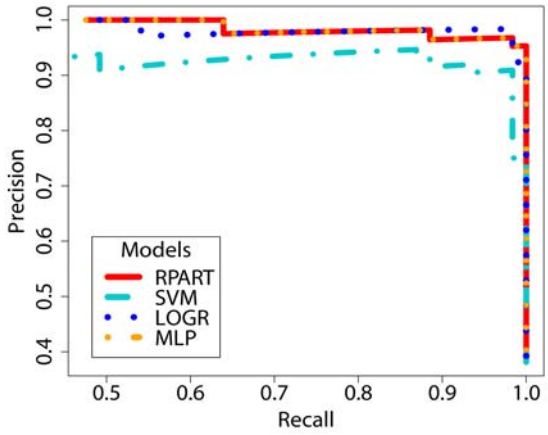

Fig. 11: Precision, F1, and Area under the ROC curve

models as it express and plots both of type 1 and type 2 errors (α and β respectively). Despite MLP and RPART models showing an identical performance on precision/recall curve. However, RPART model showing slightly higher type 1 error, which makes MLP predictive model superior over this curve. This what we could not see clearly over the ROC curve.

IV. Discussion and Conclusion

To investigate ability of machine-learning, supported by predictive analysis, for early predication of CKD, an experimental procedure has undertaken in this study, considering a dataset collected from Apollo Hospitals—India, containing 400 instances. Two class labels used as targets in the study (i.e. patients with CKD and healthy individuals), over which four machine-learning methods were simulated. The classification and regression tree, i.e. RPART model, showing considerably good result. It uses the ratio of information gain for splitting criterion, where the optimal spilt would decrease impurity of resulting subsets. In this study, RPART stopping criterion of splitting was five, which means that next split will not occur

TABLE II: Predictive power of MLs

ML methods	TPR	TNR	PPV	ACC	F1 score	AUC	Overall error	Time(milliseconds)
RPART	0.9339	1.000	1.000	0.956	0.965	0.982	4%	10
SVM	0.9892	0.8955	0.9292	0.950	0.958	0.973	5%	30
LOGR	0.9897	0.9677	0.9797	0.981	0.984	0.994	2%	50
MLP	0.9897	0.9677	0.9797	0.981	0.984	0.995	2%	40

unless there are five instances in a leaf node. Furthermore, we have identified an equal prior probability for class attribute. RPART predictive model for early predication of CKD consists of seven terminal nodes.

In this experimental procedure, we have applied two black-box models for early prediction of CKD, i.e. SVM and MLP models. A 7-7-1 MLP neural network architecture shows the highest AUC of 0.995 and TPR of 0.9897. The output of MLP model is somewhat difficult to express in comparison with RPART or LOGR models. LOGR model allows a simple calculation of the probability of prediction using the regression equation. When MLP requires connection weights of 71 to predict CKD in early stage, LOGR requires only 7 coefficients to predict the same output. MLP model uses a computational intensive back propagation algorithm to adjust connection weights and identify the ideal set of weights and bias values to predict CKD, while minimising error rate. SVM model, on the other hand, is one of the binary classification models using kernel-based learning methods. For this research, a degree 2 polynomial kernel has been employed for predication of CKD in early stages. With 16 support vectors, MLP creates a decision boundary in features space, wish is also known as hyper-plane, the ideal decision boundary should maximise the margin between healthy individuals and patients with CKD for an optimal predication.

Results showed that the highest AUC and TPR have achieved by MLP model, while the highest TNR of 1.00 has obtained by RPART model. Although RPART model can be interpreted as sets of decision rules. However, the main down-side of RPART model is considering a single parameter at each splitting process, whereas takes into account combination of parameters could lead to better predication of CKD. Moreover, MPL model showing lowest type 1 error, which makes her the best-performed predictive model. This is mainly because MLP is adaptive to handle complicated predictions. Hidden nodes enable neural network to model complex relationships between parameters as well as handle nonlinearity in data. Totally, results illustrate that machine learning represents an encouraging and viable approach for early prediction of CKD.

V. Limitations

No information about any kind of medications has been collected with this data. Value of some parameters maybe affected by prescribed drugs. For instance, patients may be prescribed a drug to control blood pressure. In this case, healthy as well as patients with CKD will record approximate values of blood pressure, which in turn weakens predictive power of blood pressure parameter.

VI. Acknowledgement

We would like to show our gratitude to the University of Anbar—Iraq for supporting this study via Computer Centre, who provided expertise that greatly assisted this research, and we thank 3 anonymous reviewers for their insights.

References

[1] V. Jha, et al., Chronic Kidney Disease: global dimension and perspectives, The Lancet, vol. 382, no. 9888, 2013, pp. 260-272; DOI https://doi.org/10.1016/S0140-6736(13)60687-X.

[2] R. Ruiz-Arenas, et al., A Summary of Worldwide National Activities in Chronic Kidney Disease (CKD) Testing, The electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine, vol. 28, no. 4, 2017, pp. 302-314.

[3] J.A. Vassalotti, et al., Practical Approach to Detection and Management of Chronic Kidney Disease for the Primary Care Clinician, The American Journal of Medicine, vol. 129, no. 2, 2016, pp. 153-162.e157; DOI https://doi.org/10.1016/j.amjmed.2015.08.025.

[4] Scottish Intercollegiate Guidelines Network (SIGN), Diagnosis and management of chronic kidney disease: A national clinical guideline, 2008.

[5] M. Kerr, et al., Estimating the financial cost of chronic kidney disease to the NHS in England, Nephrology Dialysis Transplantation, vol. 27, no. Suppl 3, 2012, pp. iii73-iii80; DOI 10.1093/ndt/gfs269.

[6] O. Dmitrieva, et al., Association of anaemia in primary care patients with chronic kidney disease, BMC Nephrology, vol. 14, 2013; DOI 10.1186/1471-2369-14-24.

[7] L.J. Rubini, Chronic kidney disease, The UCI machine-learning repository, 2015.

[8] C.C. Aggarwal, Linear Models for Outlier Detection, Outlier Analysis, Springer International Publishing, 2017, pp. 65-110.

[9] A.J. Aljaaf, Intelligent Systems Approach for Classification and Management of Patients with Headache, Phd thesis, Computer Science, Liverpool John Moores University, Liverpool, 2017.

[10] D. Ghosh and A. Vogt, Outliers: An Evaluation of Methodologies, Proc. Section on Survey Research Methods - Joint Statistical Meetings, American Statistical Association, 2012, pp. 3455-3460.

[11] C.T. Tran, et al., Multiple Imputation and Ensemble Learning for Classification with Incomplete Data, Springer International Publishing, 2017, pp. 401-415.

[12] J.L. Schafer, Multiple imputation: a primer, Statistical Methods in Medical Research, vol. 8, no. 1, 1999, pp. 3-15; DOI 10.1177/096228029900800102.

[13] J.P. Reiter and T.E. Raghunathan, The Multiple Adaptations of Multiple Imputation, Journal of the American Statistical Association, vol. 102, no. 480, 2007, pp. 1462-1471; DOI 10.1198/016214507000000932.

[14] L. PiÃl'roni, et al., Did creatinine standardization give benefits to the evaluation of glomerular filtration rate?, The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine, vol. 28, no. 4, 2017, pp. 251-257.

[15] C. Higgins, Urea and the clinical value of measuring blood urea concentration, https://acutecaretesting.org, 2016.

[16] A.S. Levey and L.A. Inker, Assessment of Glomerular Filtration Rate in Health and Disease: A State of the Art Review, Clinical Pharmacology and Therapeutics, vol. 102, no. 3, 2017, pp. 405-419; DOI 10.1002/cpt.729.

[17] E.-H. Nah, et al., Comparison of Urine Albumin-to-Creatinine Ratio (ACR) Between ACR Strip Test and Quantitative Test in Prediabetes and Diabetes, Annals of Laboratory Medicine, vol. 37, no. 1, 2017, pp. 28-33; DOI 10.3343/alm.2017.37.1.28.

[18] T.H. Jafar, et al., Progression of chronic kidney disease: The role of blood pressure control, proteinuria, and angiotensin-converting enzyme inhibition: a patient-level meta-analysis, Annals of Internal Medicine, vol. 139, no. 4, 2003, pp. 244-252; DOI 10.7326/0003-4819-139-4-200308190-00006.
[19] J. Jin, et al., Data Normalization to Accelerate Training for Linear Neural Net to Predict Tropical Cyclone Tracks, Mathematical Problems in Engineering, vol. 2015, 2015, pp. 8; DOI 10.1155/2015/931629.
[20] Z. Mustaffa and Y. Yusof, A comparison of normalisation techniques in predicting dengue outbreak, Proc. International conference on business and economics research, IACSIT Press, 2011, pp. 345-349.
[21] Al-Jumeily D, Iram S, Vialatte F-B, Fergus P, Hussain A. A Novel Method of Early Diagnosis of Alzheimer's Disease Based on EEG Signals. The Scientific World Journal, 2015; DOI 10.1155/2015/931387.