IEEE Journal of Translational Engineering in Health and Medicine

Received 13 December 2020; revised 21 February 2021; accepted 12 April 2021.
Date of publication 15 April 2021; date of current version 26 April 2021.

Digital Object Identifier 10.1109/JTEHM.2021.3073629

Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening

MD. RASHED-AL-MAHFUZ[1], ABEDUL HAQUE[2], AKM AZAD[3], SALEM A. ALYAMI[4], JULIAN M. W. QUINN[5], AND MOHAMMAD ALI MONI[6]

[1]Department of Computer Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[2]Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[3]iThree Institute, University of Technology Sydney, NSW 2007, Australia
[4]Department of Mathematics and Statistics, Imam Muhammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia
[5]Bone Biology Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
[6]WHO Collaborating Centre of eHealth, School of Public Health and Community Medicine, University of New South Wales, Sydney, NSW 2052, Australia

CORRESPONDING AUTHOR: M. A. MONI (m.moni@unsw.edu.au)

ABSTRACT Objective: Chronic kidney disease (CKD) is a major public health concern worldwide. High costs of late-stage diagnosis and insufficient testing facilities can contribute to high morbidity and mortality rates in CKD patients, particularly in less developed countries. Thus, early diagnosis aided by vital parameter analytics using affordable computer-aided diagnosis could not only reduce diagnosis costs but improve patient management and outcomes. Methods: In this study, we developed machine learning models using selective key pathological categories to identify clinical test attributes that will aid in accurate early diagnosis of CKD. Such an approach will save time and costs for diagnostic screening. We have also evaluated the performance of several classifiers with k-fold cross-validation on optimized datasets derived using these selected clinical test attributes. Results: Our results suggest that the optimized datasets with important attributes perform well in diagnosis of CKD using our proposed machine learning models. Furthermore, we evaluated clinical test attributes based on urine and blood tests along with clinical parameters that have low costs of acquisition. The predictive models with the optimized and pathologically categorized attributes set yielded high levels of CKD diagnosis accuracy with random forest (RF) classifier being the best performing. Conclusions: Our machine learning approach has yielded effective predictive analytics for CKD screening which can be developed as a resource to facilitate improved CKD screening for enhanced and timely treatment plans.

INDEX TERMS Attribute selection, chronic kidney disease (CKD), computer-aided diagnosis, explainable AI, machine learning (ML).

I. INTRODUCTION

Chronic kidney disease (CKD) is a non-communicable disease that causes large numbers of deaths worldwide, which is exacerbated by the difficulties and high costs needed for proper detection and diagnosis [1]–[5]. CKD patients show serious dysfunctions of the nervous and immune systems that severely affect their quality of life and affects many of their daily activities. Kidney failure can result in the late stages of this disease, necessitating dialysis and transplant therapies. However, adverse CKD outcomes can be reduced or prevented by early diagnosis and appropriate treatment [6], [7]. Typically early stage CKD causes little or no overt disease symptoms that cause patients to seek treatment, which makes treatment later less effective [8]. Thus better, cheaper and more effective screening tools would reduce the burden of disease simply by identifying at risk individuals at an early stage. This is made more difficult to achieve by the fact that underlying pathogenic mechanisms for CKD are largely unclear. Thus, many remain undiagnosed for a long period [9]. The National Health and Nutrition Examination Survey 2003-2004 reported that only around 5 percent of CKD patients with stage 1 or 2 and less than 10 percent patients with stage 3 have been diagnosed with CKD, and only 45% of CKD patients in their stage 4 were known about

their condition and physical symptoms [10], [11]. Therefore, early detection of CKD is crucial to enable initiation of treatment to prevent and delay CKD progression, and slow the development towards advanced stage complications. However, there is lack of proper low cost knowledge-based tools, which is particularly a point of concern for healthcare in the least developed countries. Easy and low cost CKD diagnosis would enable routine testing of kidney function and would increase prevention of late stage disease.

Traditionally there have several biomarkers and computational methods used in the diagnosis and measurement of CKD severity. The most commonly used for kidney function is the glomerular filtration rate (GFR) [12], which is, however, not sufficient to diagnose CKD [13]. GFR measurements require creatinine measurement (or creatinine clearance tests) and urine albumin tests which are not economical and practical enough for routine CKD disease screening [14]. A high GFR can be seen for some patients with cardiovascular disease and diabetes, in which case high GFR levels may mask the presence of actual CKD [13]. Furthermore, inadequate numbers of experienced nephrologists and lack of imaging and biopsy services in less developed countries means that not all patients with CKD are properly tested in a timely manner. Therefore, computer-aided automated, accurate, convenient, low-cost CKD detection technique could enhance early diagnosis and intervention.

Machine learning (ML) techniques can be used to facilitate medical diagnosis of CKD, indeed several studies have already used these techniques to improve clinical prediction in kidney-related disease and have reported an improved classification accuracy [15]–[21]. However, none of these studies identified the most important predictive attributes needed to improve diagnosis. If identified in CKD patients, such attributes could be used for computer-aided CKD screening and diagnostic tests. However, using UCI-CKD data [22], only a few studies [23]–[25] have attempted to identify those important attributes. Nishanth et. al. [23] investigated the reasons why (i.e., the underlying mechanisms for) the attributes they identified being able to improve predictions for GFR. This study used common spatial patterns (CSP) and linear discrimination analysis (LDA) to identify important features; they found that measurements of serum hemoglobin, albumin, specific gravity, biomarkers for hypertension, and biomarkers for diabetes mellitus, together with serum creatinine were important features that improved predictions. Wrapper approach was used [24] to select these important features for CKD diagnosis; this reported the important attributes were specific gravity, albumin, red blood cell numbers, pus cell clumps, serum creatinine, sodium, hemoglobin, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. Another study [25] proposed a correlation-based feature subset selection method; they investigated specific gravity, albumin, serum creatinine, hemoglobin, packed cell volume, white blood cell count, red blood cell count, and hypertension, finding these as the most significant in the detection of CKD. Another study [26] used only one CKD attribute collected from year-long temporal data, using electronic health records (EHRs). However, new patients without EHRs would not be able to use this approach. However, all these above studies presented results and selected attributes from black-box nature classification approach to model construction and the lack of interpretation of the decision in the diagnostic model can lead to adverse or even life-threatening consequences. Moreover, there is lack of proper rationales for selecting particular attributes for model decision making in the existing studies. Thus, an active area of research is in the construction of interpretable ML models to use computer-aided diagnostic systems, which would allow clinicians to better evaluate model decisions and discern the role of particular model attributes used in decision making.

In this study, we have used ML models for the early diagnosis of CKD, and not only have attempted identification of key predictive features but have also used model interpretation techniques in the attribute selection, namely SHapley Additive exPlanations (SHAP). The main objective of this study is to reduce the number of attributes to a minimum in order to find an optimal set useful for clinical testing and to achieve a high CKD detection accuracy with them. This study therefore determines the applicability of ML-based models to the diagnosis of CKD using urine or blood test with the smallest number of additional attributes.

II. DATA AND METHODS

The present work was accomplished in several stages: data collection, pre-processing, model training, important attributes selection, and evaluation of the models and selected attributes, which are summarised in a schematic diagram in Fig. 1. The labeled data was collected from hospital-based sources [for details see subsection below] [22]. In the pre-processing stage, the database was analysed and various techniques applied to transform the data into the proper structure, with missing value imputations. After completing the pre-processing, we split the data into training and test sets. The training data set was used to train a number of ML classifier models; these included random forest (RF), gradient boosting (GB), XGBoost (XGB), Logistic Regression (LR), and support vector machine (SVM) models. Subsequently the SHAP technique was employed to interpret the model decisions and identify the important features that contributed most to the classification process. The selected highest importance attributes were then used to form new reduced datasets with which the classifiers were trained and tested to determine whether those attributes were sufficient to build an ML-based computer-aided CKD diagnosis system that shows high classification accuracy. This approach will allow users to predict and diagnose their possible CKD cases using the minimum necessary number of clinical tests, which will reduce cost and resources.

A. DATA

This CKD dataset used in this study was released by Apollo Hospitals, Tamil Nadu, India, in July 2015 and is available in

IEEE Journal of Translational Engineering in Health and Medicine

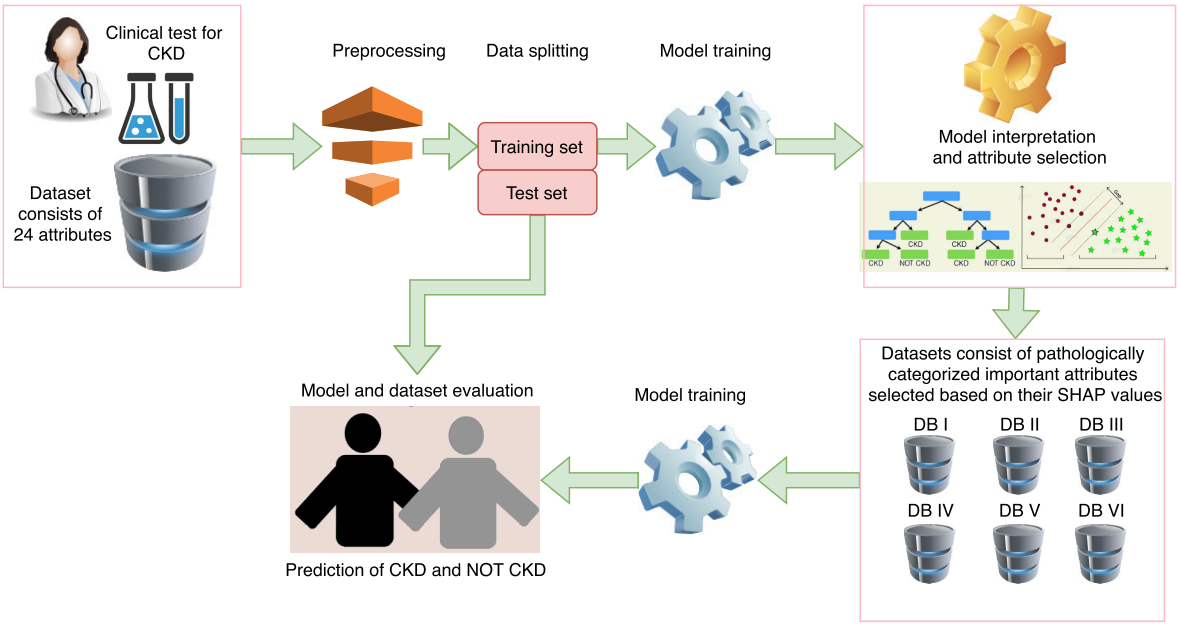

FIGURE 1. Schematic diagram of the overall workflow.

the UCI machine learning repository [22]. The dataset consists of samples collected from 400 patients with 24 attributes (13 nominal and 11 numerical attributes). The 24 attributes are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. The description of the dataset is given in Table 1. Based on these available clinical attributes 250 cases were classified as CKD and the remaining 150 were classified as non-CKD.

B. PRE-PROCESSING

We have pre-processed the unrefined medical data by removing the missing values to enhance prediction capabilities. We have also conducted data-transformation to make them useful for the machine learning models, which are limited to process non-numerical data. The non-numerical data in the dataset are in the form of 'present', 'not present', 'normal', 'abnormal', 'yes', 'no', 'good', and 'poor'. The non-numerical data are identified and transformed into numbers. The 'normal', 'present', 'yes', and 'good' values for nominal attributes are replaced by '1' and 'abnormal', 'notpresent', 'no', and 'poor' values are replaced by '0'.

Missing values are concomitant to real-world data. Ignorance of the record that contains the missing value is the simplest form of a solution, which is a less-desirable practice, especially for small dataset. In the whole data set, there are only 9.70% missing values. We used various imputation algorithms, arithmetic mean and mode imputations that show good performance in some studies [28], [29], to solve the missing value problem rather than merely removing the records. The numerical attributes were arithmetic mean imputed where the missing values are replaced with the represented mean value of that attribute. In the case of nominal attributes, mode imputation is performed where the missing values are replaced with the most frequently occurred value of that attribute. After pre-processing, the data distribution is transformed, which is depicted in Fig. 2.

C. MACHINE LEARNING MODELS

1) RANDOM FOREST (RF)

Random forest (RF) is an ensemble learning technique that has been proven to be very effective and powerful classifier [30]. RF consists of a combination of many decision trees where each tree is trained on a randomly selected feature vector from the training data set. For a new test sample, each tree of the forest classifies it individually and yields a certain classification result. The RF decides the predicted class of the test data depending on the majority of votes over all the trees in the network.

2) GRADIENT BOOSTING (GB)

The gradient boosting algorithm [31] produces a prediction model that consists of an ensemble of weak prediction models. This means a collection of individual models results in a final model. The model is built in a stage-wise fashion. The individual models have poor prediction power and suffer over-fitting problems, but the ensemble of these models provides improved results. The individual models in the ensemble are not built on completely random subsets of training data but by putting more weights on the wrong predicted

TABLE 1. Details information of the attributes of the UCI CKD dataset [22], [27].

Attributes	Description	Type of test	Attribute Type	Attribute values
age	Age	-	numeric	years
bp	Blood Pressure	-	numeric	mm/Hg
sg	Specific Gravity	Urine	numeric	1.005, 1.010, 1.015, 1.020, 1.025
al	Albumin	Urine	numeric	0, 1, 2, 3, 4, 5
su	Sugar	Urine	numeric	0, 1, 2, 3, 4, 5
rbc	Red Blood Cells	Urine	nominal	normal, abnormal
pc	Pus Cell	Urine	nominal	normal, abnormal
pcc	Pus Cell Clumps	Urine	nominal	present, notpresent
ba	Bacteria	Urine	nominal	present, notpresent
bgr	Blood Glucose Random	Blood	numeric	mgs/dl
bu	Blood Urea	Blood	numeric	mgs/dl
sc	Serum Creatinine	Blood	numeric	mgs/dl
sod	Sodium	Blood	numeric	mEq/l
pot	potassium	Blood	numeric	mEq/l
hemo	Hemoglobin	Blood	numeric	gms
pcv	Packed Cell Volume	Blood	numeric	-
wbcc	White Blood Cell Count	Blood	numeric	cells/cumm
rbcc	Red Blood Cell Count	Blood	numeric	millions/cmm
htn	Hypertension	-	numeric	yes, no
dm	Diabetes Mellitus	-	numeric	yes, no
cad	Coronary Artery Disease	-	nominal	yes, no
appet	Appetite	-	nominal	good, poor
pe	pedal Edema	-	nominal	yes, no
ane	Anemia	-	nominal	yes, no
class	Class	-	nominal	ckd, notckd

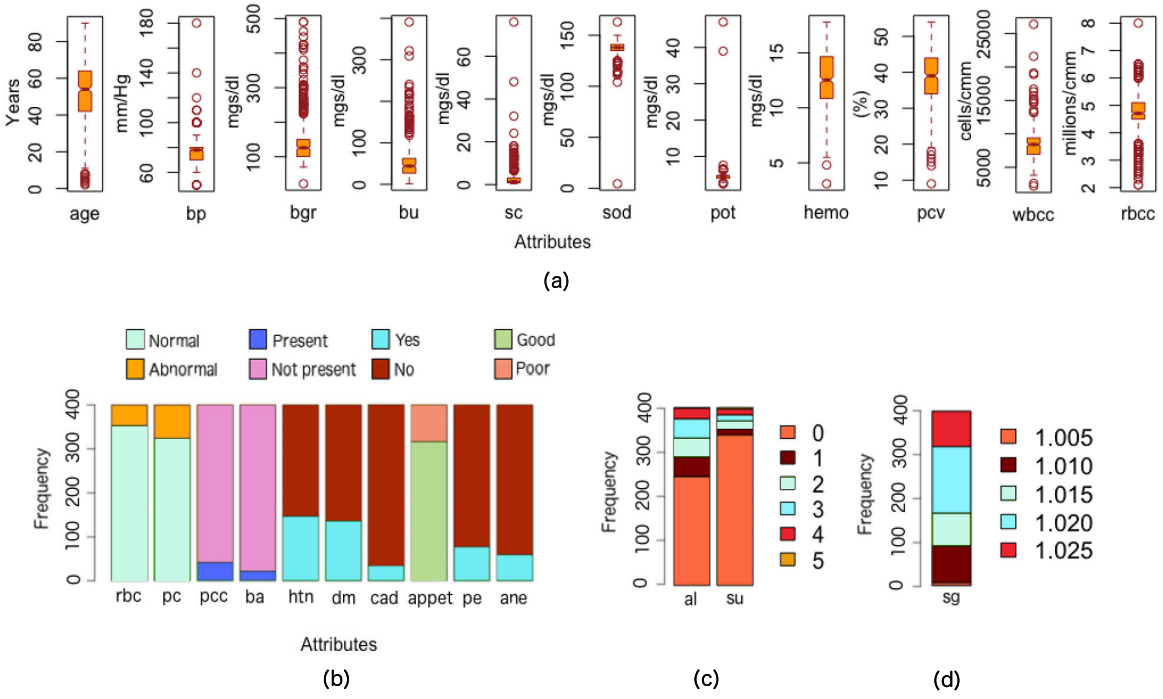

FIGURE 2. Sample frequency distribution of clinical test attributes. (a) sample frequency distribution (box plot) of numeric attributes, (b) sample frequency distribution (bar plot) of two-class nominal attributes, (c) sample frequency distribution (bar plot) of six-class nominal attributes, and (d) sample frequency distribution (bar plot) of five-class 'sg' nominal attributes.

samples; in other words, instances that are hard to predict will be more focused during model training by taking into account the past mistakes. In Gradient Boosting, the predictions of the models are combined. For this reason, the boosted model predictions are optimized instead of optimizing the model parameters directly.

3) EXTREME GRADIENT BOOSTING (XGB)

Extreme Gradient Boosting (XGB) [32] is a scalable implementation of gradient boosting and finds the best tree model. It computes second-order gradients to get more information about the direction of gradients and minimize the loss function. Unlike the base model, such as a decision tree

that uses loss function as a proxy for minimizing the overall model's cost, XGB uses second-order derivative as an approximation. To improve model generalization, XGB uses advanced L1 and L2 regularization techniques.

4) LOGISTIC REGRESSION (LR)

Logistic regression (LR) [33] is a widely used classifier for binary classification problems. It finds a function that predicts the outcome for a binary dependent variable from one or more independent variables. The sigmoid function plays a crucial role in the logistic regression classifier. The sigmoid function provides the output as a number between 0 and 1. A threshold is used to consider the output as to belong to class 1 or class 0. An input sample is considered to belong to class 1 if the output greater than 0.5, otherwise the classifier considers it belongs to class 0.

5) SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) [34], [35]-a widely used supervised machine learning method that is capable of identifying subtle patterns in noisy and complex datasets and used for binary classification. SVM is developed based on statistical learning theory. It uses several kernel functions to project non-linearly separable samples in lower-dimensional space onto another higher dimensional space.

D. MODEL INTERPRETATION FOR FEATURE SELECTION

Classification models map a test instance to output and most of the cases provide a single metric, such as classification accuracy. However, this is not a complete description of why the model made this correct prediction. Sometimes it is useful to determine how the prediction was made and the role of the individual features in making this decision. Model interpretability can let us know the feature importance of the model. We can also understand predictions from methods by attributing importance values to each input feature. To understand the classifier's overall behaviour, these important values can be computed either for a single prediction, or an entire dataset. We calculated feature importance using the training dataset.

SHapley Additive exPlanations (SHAP) technique, proposed by Lundberg and Lee [36], has been shown to be effective for identifying important features in the dataset. SHAP uses popular game theory rules [37] and local explanations methods [38], and is able to estimate the degree of contribution of each feature to the overall decision making ability of the model. Given a model with a set of all the input features, N, to predict output $f(N)$, SHAP values are calculated using several axioms to allocate the contribution of each feature using the following equation:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(K - |S| - 1)!}{K!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

where ϕ_i is the feature importance of ith attribute to make the output decision of the model and it is assigned based on their marginal contribution [39], K is the number of input features, and S is the set of non-zero indexes in z'.

An additive feature attribution method is used to define a linear function, h, of binary variable as Equation 2 where $z' \in \{0, 1\}^K$ equals to 1 when a feature is observed, otherwise it equals to 0 [36].

$$h(z') = \phi_i + \sum_{i=1}^{K} \phi_i z'_i \quad (2)$$

where $\phi_i \in \mathbb{R}$

Here we use Tree SHAP, the fast SHAP value computation method for RF, GB, and XGb, and linear explainer for LR, and SVM models [40] to identify an important feature of the training dataset.

III. RESULTS

A. MODEL TRAINING AND FEATURES SELECTION

Among the ten-folds-split data nine folds were used to train all the classifier models. Then the remaining 10th-fold data was used to evaluate the classifiers. Thus we had 360 samples for the training dataset and 40 samples for the test dataset. The training dataset contained 90% of the total samples, and the testing dataset contained 10% of the total samples. There were 225 CKD cases and 135 non-CKD cases in the training nine folds dataset, and in the remaining test one fold dataset, there were 25 CKD cases and 15 non-CKD cases. Six evaluation matrices were measured for each classifier to observe and compare the performances of the models. According to the evaluations matrices RF classifier showed the highest predictive performance for the test datasets with an average classification accuracy of 99.50%, sensitivity of 98.75%, specificity of 100%, precision of 100%, F1 score of 99.35%, and AUC of 99.38%, shown in Fig-3(a). The GB and XGB classifiers showed almost similar classification performances. The GB classifier obtained classification accuracy of 99.00%, sensitivity of 98.79%, specificity of 99.21%, precision of 98.66%, F1 score of 98.71%, and AUC of 99.00%, and the XGB classifier obtained classification accuracy of 98.75%, sensitivity of 98.20%, specificity of 99.25%, precision of 98.62%, F1 score of 98.34%, and AUC of 98.72%. The classification performances for LR and SVM are relatively lower than the performances of the RF, GB, and XGB classifiers, shown in Fig. 3(b).

Global feature importance was computed in terms of SHAP values. After training the classifiers with the training data sets including all the 24 attributes, SHAP values of every attribute were calculated to understand the feature importance of the classifiers. Fig. 4(a) shows the feature importance values obtained from RF classifier trained on fold-1 training data set. The magnitude of the feature values is color-coded from black to copper for the feature values from low to high, respectively. The features are vertically sorted according to their average impact on the predictions. From Fig. 4(a), it is shown that the first 13 features are considered as the most important feature. The feature importance values were calculated for

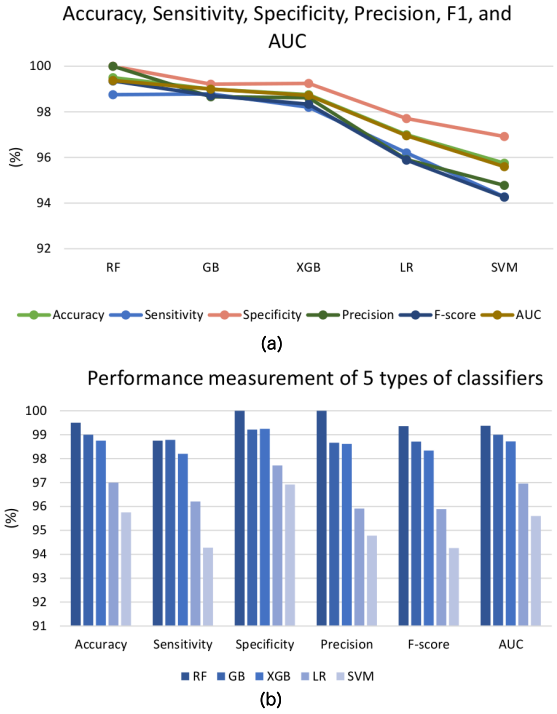

FIGURE 3. Performance of the classifiers using all the attributes. (a) Various model evaluation metrics for CKD classification. (b) Calculated classification accuracy for RF, GB, XGB, LR, and SVM classifiers.

classifiers using all the training data sets. Fig. 4(b) depicts the normalized mean SHAP value magnitudes obtained from each of the RF, GB, and XGB models using all the training set. The RF, GB, and XGB classifiers were considered based on their better performance than the LR and SVM classifiers. According to Fig. 4(b), hemo, sg, sc, al, pcv, rbcc, htn, bgr, dm, age, sod, bu, and bp attributes had the highest impact on the RF, GB, and XGB classifier's decision making on whether the instance is CKD or NOT CKD. The identified important features were then used to reduce the dimensionality of the dataset and minimize the costs and efforts of the diagnosis of CKD with perfect accuracy. The selected features had been used to produce a new lower-dimensional dataset than the original dataset. Hence, after discarding the less important features, the SHAP-based feature identification method reduced the dataset dimension to 13 attributes, which were further subjected to build six different datasets based on the different pathological tests in order to observe the predictive performances of ML models on the different combinations of the attribute sets.

B. INTERPRETATION OF THE IMPACT OF ATTRIBUTION INTERACTIONS ON CKD PREDICTION

The interactions of the attributes on prediction offer more in-depth insight into models' decision making capabilities. The SHAP dependence contributions between attributes were observed to analyse the impact of attribute interaction on the model's decision making capabilities.

The dependence contribution plots for the hemoglobin and specific gravity attributes is shown in Fig. 5(a), where the x-axis represents hemoglobin level and the y-axis represents SHAP value for hemoglobin, and the color bar represents the values associated with specific gravity. Thus each dot in the plot represents each value of the hemoglobin attribute in the training dataset. Lower values of both hemoglobin and specific gravity made a higher impact on CKD prediction, though the interaction of the lower value of hemoglobin with some higher value of specific gravity produced higher SHAP value for hemoglobin. From the figure, it is clearly observed that the highest impacts on chronic kidney disease detection were achieved when hemoglobin was less than 13 gms and specific gravity was equal to or less than 1.015. The value of specific gravity less than or equal to 1.015 also obtained higher SHAP values while interacting with other attributes.

Fig. 5(b) shows that the interaction of very high value of serum creatinine and very low value of hemoglobin resulted in higher SHAP value for hemoglobin. Higher albumin and lower hemoglobin interaction caused higher SHAP values for hemoglobin. The dependence contribution plots for the hemoglobin and packed cell volume attributes are shown in Fig. 5(d). The lower values of both hemoglobin and pcv were responsible for achieving higher SHAP values of hemoglobin. The highest impacts on chronic kidney disease detection are observed when hemoglobin is less than 13 gms and packed cell volume is less than 40. The interaction of the lower value of hemoglobin with lower red blood cell count values enhanced the chance of the CKD prediction (Fig. 5(e)). Fig. 5(f) shows the impact of the interaction of hemoglobin and hypertension on CKD prediction. Higher hypertension was responsible for achieving the higher SHAP values of hemoglobin. The higher values of blood glucose random, diabetes mellitus, blood urea, and blood pressure interacted with lower value of hemoglobin made a higher impact on CKD prediction, shown in Fig. 5(g, h, k, l). All the ages were interacted with lower hemoglobin to produce higher SHAP value of hemoglobin, shown in Fig. 5(i), but relatively larger ages mostly interacted with lower hemoglobin in the model for CKD prediction. Lower value of sodium interacted with lower value of hemoglobin in the model to make the decision the attributes were from a CKD patient, shown in Fig. 5(j), although a small number of lower sodium interacted with higher hemoglobin to predict NOT CKD.

C. REDUCED DATASET AND MODELS EVALUATION

Based on the calculated SHAP values of the attributes with the RF, GB, and XGB models, 13 attributes were selected to reduce the dataset. The selected attributes are shown in Table 2. We created six different datasets by taking attributes from the selected attributes and based on test pathologies, shown in Table 3. Dataset 'DB-I' consists of all the selected 13 attributes. 'DB-II' dataset consists of attributes collected from blood and other pathological tests, including hemo, sc, pcv, rbcc, bgr, sod, and bu from the blood test and htn, dm, age, and bp from other tests attributes.

IEEE Journal of Translational Engineering in Health and Medicine

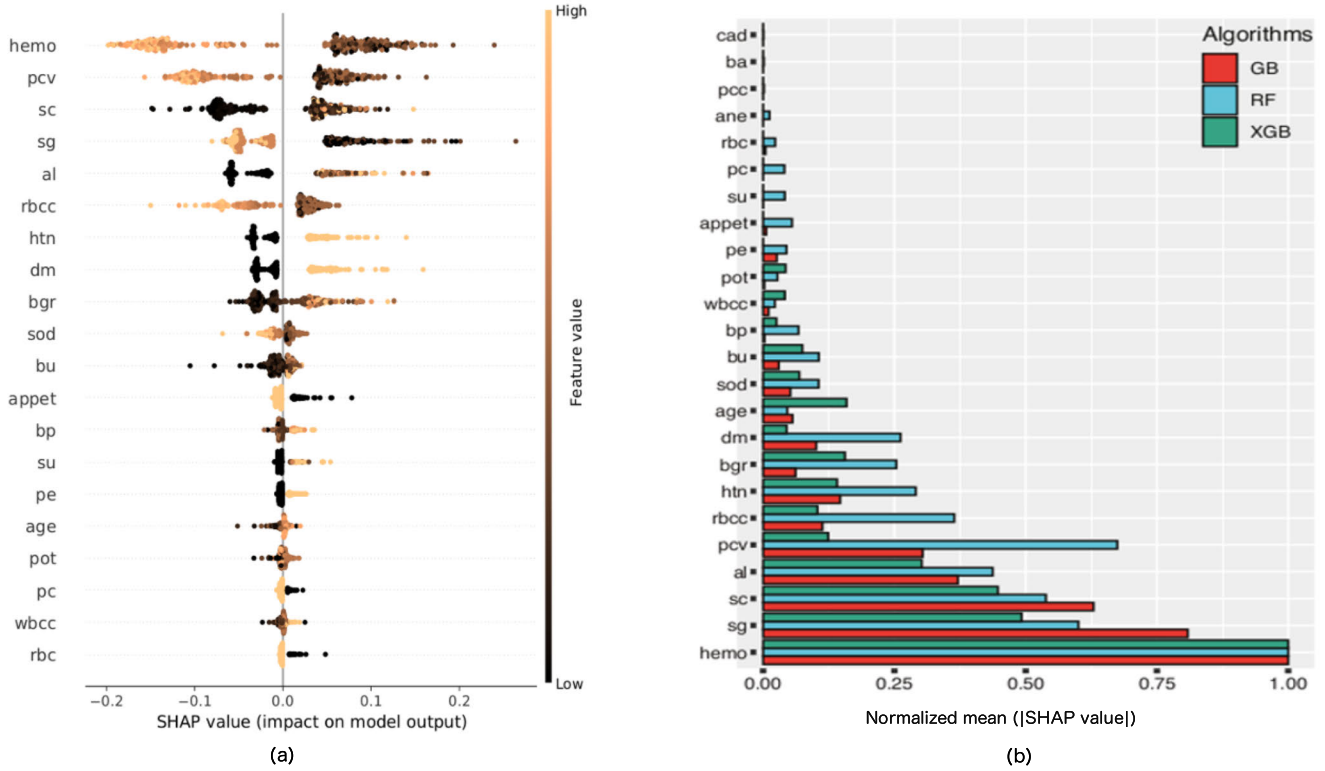

FIGURE 4. SHAP plots (a) SHAP plot of the top most 20 attributes. The SHAP values were calculated when the random forest (RF) model was trained for a single fold training dataset where each dot corresponds to an instance from the training data. Each value is color coded, dark black represents the lower value and light color represents the higher value of the attributes. (b) The bar plot represents normalized mean absolute SHAP value across all the folds for the RF, GB, and XGB model training.

TABLE 2. Lists of important attributes categorized according to pathology.

Blood test attributes	Urine test attributes	Other test attributes
hemo	sg	htn
sc	al	dm
pcv		age
rbcc		bp
bgr		
sod		
bu		

TABLE 3. Datasets consists of selected clinical test attributes used for CKD diagnosis.

Database	Pathology	Attributes
DB I	Blood + Urine + other	hemo, sc, pcv, rbcc, bgr, sod, bu, sg, al, htn, dm, age, bp
DB II	Blood + other	hemo, sc, pcv, rbcc, bgr, sod, bu, htn, dm, age, bp
DB III	Urine + other	sg, al, htn, dm, age, bp
DB IV	Blood	hemo, sc, pcv, rbcc, bgr, sod, bu
DB V	Urine	sg, al
DB VI	other	htn, dm, age, bp

Database 'DB-III' consists of urine test attributes, sg and al, and 4 other tests attributes. 'DB-IV' includes blood test attributes only, 'DB-V' includes urine test attributes only, and 'DB-VI' includes other test attributes excluding both blood and urine test attributes.

All the models were trained using the samples that were previously used to train the models for SHAP calculation. The models were tested using the test datasets that had not been previously introduced to the models. The datasets and models were evaluated for classification accuracy, sensitivity, specificity, precision, F-score, and the area under the curve (AUC), and results obtained are shown in Fig. 6.

The RF model presented the highest classification accuracy of 99.00% among other classifiers with the 'DB-I' dataset, whereas the accuracy of 98.25%, 98.50%, 97.25%, and 97.75% were obtained using GB, XGB, LR, and SVM classifiers, respectively. The 'DB-II' dataset consisted of selected blood and other test attributes including 11 attributes (7 blood and 4 other). RF gave the highest CKD classification performances with the accuracy of 97.75%, sensitivity of 96.12%, specificity of 98.82%, precision of 97.94%, F-score of 96.88%, and AUC of 97.47%. The GB classifier obtained an accuracy of 96.75%, sensitivity

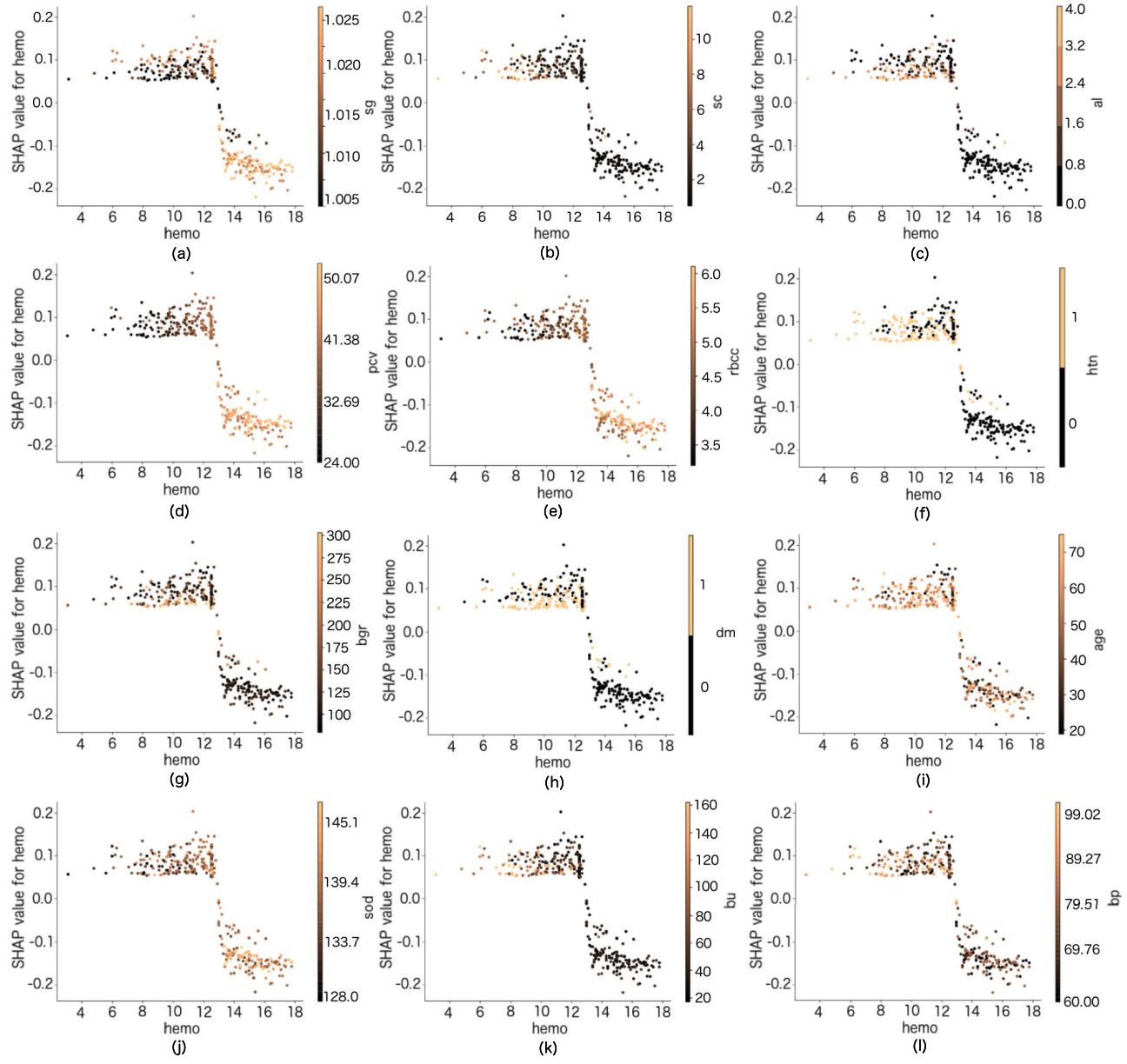

FIGURE 5. **Attributes dependence plots for the interaction of hemoglobin and other attributes. X-axis represents hemoglobin level and Y-axis represents the SHAP value of hemoglobin in the RF model. Copper color in the color bars represents higher values, and dark color present lower values of the attributes. (a-l) Interaction effects with specific gravity, serum creatinine, albumin, packed cell volume, red blood cell count, hypertension, blood glucose random, diabetes mellitus, age, sodium, blood urea, and blood pressure, respectively.**

of 96.12%, specificity of 97.28%, precision of 95.43%, F-score of 95.55%, and AUC of 96.70%. The performance of XGB was a little bit higher than the GB, LR, and SVM classifiers with the classification accuracy of 97.00%. For the 'DB-III' dataset, both GB and XGB performed better than the RF classifiers with an accuracy of 97%. However, RF detected CKD with the highest specificity and precision than other classifiers. For the 'DB-IV' dataset, RF again achieved the highest classification accuracy of 97.25% than other classifiers. The models showed relatively lower classification accuracy with the 'DB-V' and the 'DB-VI' than the other datasets, where RF achieved a classification accuracy of 88.50% and 86.25% using the 'DB-V' and the 'DB-VI' datasets, respectively.

Database evaluation for the RF classifier, which showed better classification for most of the cases, is shown in Fig-7. The database DB I showed the best CKD detection performance which is almost similar to the full database including 24 attributes. However, DB II, DB III, and DB IV also showed considerable classification performances for CKD detection. The results indicate that using all 24 features did not bring additional advantages in the diagnosis process, but took time

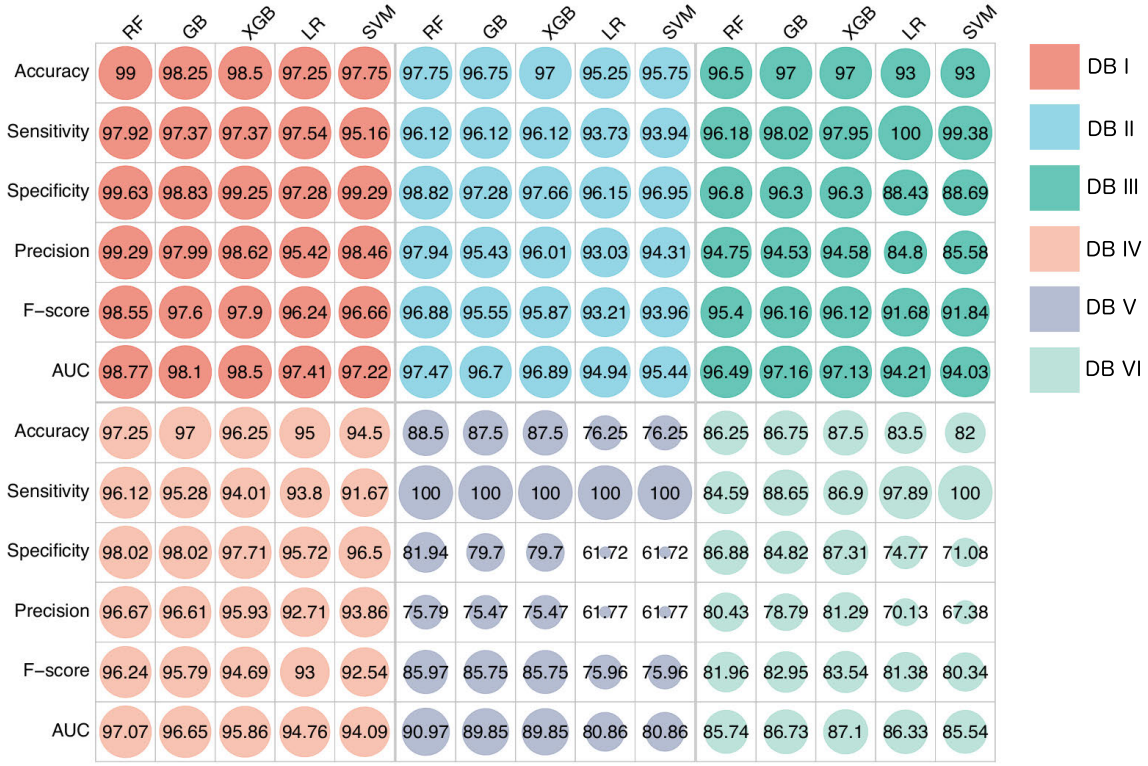

FIGURE 6. CKD detection accuracy plots for the six versions of the dataset, namely 'DB-I', 'DB-II', 'DB-III', 'DB-IV', 'DB-V', and 'DB-IV' datasets with various machine learning models. The performance metrics of specific databases are color coded in the plot.

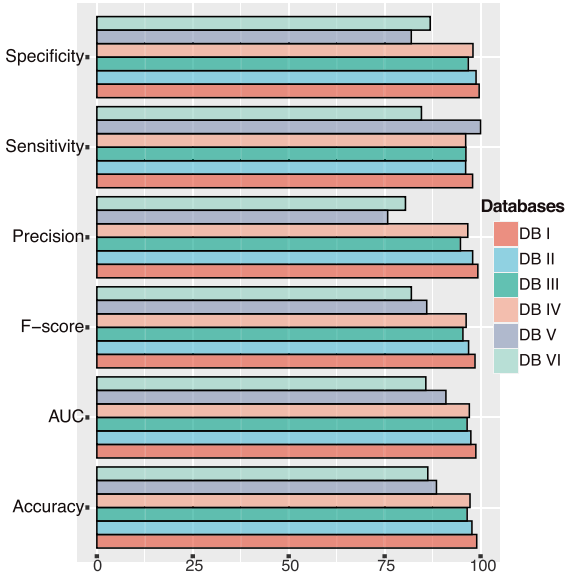

FIGURE 7. A bar chart of the CKD classification accuracy in the RF model for database evaluation. There are six datasets including various test attributes and the performance rate range is from 0 to 100%. The superior performances of the model trained with DB I were observed compared to those models trained with other datasets, in terms of specificity, precision, F-score, AUC and Accuracy, but, however, DB V trained model showed only marginally better sensitivity than that of DB I.

and efforts and extra cost. Because using the only selected important 13 attributes in DB I, 7 blood and 4 other test attributes in DB II, 2 urine and 4 other test attributes in DB III, and only 7 blood attributes in DB IV, we achieved almost similar classification accuracy.

IV. DISCUSSION

The primary objective of this study was to identify important clinical test attributes not only to enable efficient computer-aided CKD screening but also to help reduce the costs of CKD diagnosis. Results obtained using our proposed framework indicate that the ML models showed better CKD and non-CKD classification with a considerably reduced number of attributes, 13 out of 24 that were employed. Using a higher number of test attributes than necessary has a significant financial impact, which hinders the routine screening for CKD. For this reason we applied the recently developed SHAP technique to identify important attributes to the classifiers for CKD detection. We have studied five ML models with these clinical test attributes to choose the most suitable classifier that can diagnose CKD with superior accuracy using selected important attributes obtained from single clinical pathology, either urine or blood, or both. Despite using a low number of attributes the classification accuracy of the ML models were still able to give near perfect accuracy. Notably, RF gave the best results of the five ML classifier methods tested. This study of ML techniques with reduced numbers of attributes demonstrates that it is possible to diagnose CKD at a lower cost. A person tested CKD by only low-cost urine pathology and other almost free tests such

as blood pressure, hypertension, and age can be referred to further testing of CKD with more clinical attributes. Early referral of prospectively diagnosed CKD positive person, based on urine testing, encourages further examinations with blood pathology testing. The combined blood and previously collected urine and other test attributes would thus be used for automated ML-based diagnosis for proper management and treatment of CKD patients.

Recently, the interpretation of decisions made by classifiers, and transparency in the model learning process is required by medical practitioners [41], where the reliable interpretation of the decision-making process can build trust among medical personnel. Furthermore, traditional model performance measurement metrics, i.e., accuracy, sensitivity, and specificity, indicate how well or badly a model performs, but fail to interpret what were the roles of particular attributes to contribute to the classification decisions, and how the values of the attributes influence particular decisions. To increase clinician confidence in ML automated CKD diagnosis systems, we have interpreted the decision-making in terms of attribute values. We used the explainable AI-based machine learning algorithms SHAP to provide some explanation and justification for the decision.

The ML decision of detecting CKD based on the attribute properties was comparable to current medical practices. For instance, the hemoglobin attribute was given the highest importance according to the SHAP value with low values of hemoglobin associated with the occurrence of CKD. This finding is compatible with previous observations that people with early stages of CKD have low red blood cell hemoglobin content [42]. The association of low hemoglobin readings with CKD is also supported by the finding of a working group study, namely Kidney Disease: Improving Global Outcomes Anemia Work Group (KDIGO) [43]; this reported that males older than age 15 with hemoglobin levels of less than 13 g/dL and female older than 15 years with hemoglobin levels below 12 g/dL are diagnosed as having anemia, and typically have lost at least half of their kidney function. The high value of the two selected attributes, packed cell volume and specific gravity, related to their inverse relationship to CKD occurrence. The low value of packed cell volume is associated with developing CKD, a finding also consistent with existing knowledge [44]. The presence or absence of hypertension (blood pressure at unhealthy high levels) is also relevant, as hypertension has a clear positive influence on the classifier decision on CKD diagnosis consistent with its known positive correlation with CKD [45]. The SHAP-identified importance of red blood cell count (rbcc), also suggests that lower values have a definite impact on the likelihood of CKD. The higher values of both blood glucose and diabetes mellitus similarly influence CKD prediction.

The identified importance of the three model-selected attributes blood urea (bu), sodium (sod), electrolyte, albumin (al), and serum creatinine (sc) are consistent with their use in clinical guidelines such as the Kidney Disease Improved Global Outcomes KDIGO [46], the national institute for health and care excellence [47], and Kidney Disease Outcomes Quality Initiative (KDOQI) [48]; these four biomarkers are considered as useful clinical tests for CKD diagnosis. Hypertension (htn) and diabetes mellitus (dm) are themselves chronic diseases that influence the progression of CKD [47], [49], and were both identified as important for the ML models to diagnosis CKD. Age and blood pressure, two cost free attributes, were also considered important by SHAP. When we added these four clinical attributes with two selected urine test attributes (sg and al) for model training and test purpose, the CKD identification accuracy also increased. The selected blood test attributes with hypertension, and diabetes mellitus, age, and blood pressure when all used together also gave some apparent improvement CKD diagnostic accuracy compared to blood test attributes alone. We hope that the application of our ML-empowered and explainable AI-based approach can be useful not only for designing efficient and cost-effective computer-aided CKD detection tools, but also to build medical practitioner trust in these reporting tools.

V. CONCLUSION

This paper has identified a reliable method for CKD classification and attributions selection with improved simplicity and cost effectiveness. First, we trained and selected suitable classifiers, calculated the feature importance based on SHAP values, and obtained a reduced dataset based on the pathological tests and measured feature importance. Second, we trained the classifiers with these reduced data sets and evaluated them with the test datasets. The results of this analysis demonstrated that the SHAP-identified important features were consistent with the current clinical thinking. It also found that an RF classifier method provides significantly high classification accuracy with the pathologically categorized attributes sets. The proposed RF classifier and reduced test attributes can therefore be potentially applied to reduce diagnosis costs and enable better management of early treatment plans.

REFERENCES

[1] A. N. Muiru *et al.*, "The epidemiology of chronic kidney disease (CKD) in rural east Africa: A population-based study," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0229649.

[2] C. P. Wen *et al.*, "All-cause mortality attributable to chronic kidney disease: A prospective cohort study based on 462 293 adults in Taiwan," *Lancet*, vol. 371, no. 9631, pp. 2173–2182, Jun. 2008.

[3] M. A. Hossain, T. A. Asa, M. R. Rahman, and M. A. Moni, "Network-based approach to identify key candidate genes and pathways shared by thyroid cancer and chronic kidney disease," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100240.

[4] K. Brück *et al.*, "CKD prevalence varies across the European general population," *J. Amer. Soc. Nephrol.*, vol. 27, no. 7, pp. 2135–2147, 2016.

[5] *2015 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*, United States Renal Data System, Bethesda, MD, USA, 2015.

[6] G. Remuzzi, P. Ruggenenti, and N. Perico, "Chronic renal diseases: Renoprotective benefits of renin–angiotensin system inhibition," *Ann. Internal Med.*, vol. 136, no. 8, pp. 604–615, 2002.

[7] M. A. Hossain, T. A. Asa, S. M. S. Islam, M. S. Hussain, and M. A. Moni, "Identification of genetic association of thyroid cancer with parkinsons disease, osteoporosis, chronic heart failure, chronic kidney disease, type 1 diabetes and type 2 diabetes," in *Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE)*, Sep. 2019, pp. 832–837.

[8] O. J. Wouters, D. J. O'donoghue, J. Ritchie, P. G. Kanavos, and A. S. Narva, "Early chronic kidney disease: Diagnosis, management and models of care," *Nature Rev. Nephrol.*, vol. 11, no. 8, p. 491, 2015.

[9] K.-U. Eckardt *et al.*, "Autosomal dominant tubulointerstitial kidney disease: Diagnosis, classification, and management—A KDIGO consensus report," *Kidney Int.*, vol. 88, no. 4, pp. 676–683, 2015.

[10] *National Health and Nutrition Examination Survey Data, 1999–2000, 2001–2002, and 2003–2004*, U.S. Dept Health Hum. Services, Centers Disease Control Prevention, Hyattsville, MD, USA, 2007.

[11] L. C. Plantinga *et al.*, "Patient awareness of chronic kidney disease: Trends and predictors," *Arch. Internal Med.*, vol. 168, no. 20, pp. 2268–2275, 2008.

[12] A. Yadollahpour, "Applications of expert systems in management of chronic kidney disease: A review of predicting techniques," *Oriental J. Comput. Sci. Technol.*, vol. 7, no. 2, pp. 306–315, 2014.

[13] A. S. Allen, J. P. Forman, E. J. Orav, D. W. Bates, B. M. Denker, and T. D. Sequist, "Primary care management of chronic kidney disease," *J. Gen. Internal Med.*, vol. 26, no. 4, pp. 386–392, 2011.

[14] T. Fiseha, M. Kassim, and T. Yemane, "Chronic kidney disease and underdiagnosis of renal insufficiency among diabetic patients attending a hospital in southern ethiopia," *BMC Nephrol.*, vol. 15, no. 1, p. 198, Dec. 2014.

[15] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, p. 55, Apr. 2017.

[16] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in *Proc. CMBEBIH*. Singapore: Springer, 2017, pp. 589–594.

[17] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–16, Dec. 2019.

[18] Z. Chen, Z. Zhang, R. Zhu, Y. Xiang, and P. B. Harrington, "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," *Chemometric Intell. Lab. Syst.*, vol. 153, pp. 140–145, Apr. 2016.

[19] M. E. Hossain, A. Khan, M. A. Moni, and S. Uddin, "Use of electronic health data for disease prediction: A comprehensive literature review," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 745–758, Mar. 2021.

[20] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, Jun. 2019.

[21] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A comparative analysis of active learning for biomedical text mining," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 23, Mar. 2021.

[22] L. J. Rubini *et al.*, "UCI chronic kidney disease," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2015.

[23] A. Nishanth and T. Thiruvaran, "Identifying important attributes for early detection of chronic kidney disease," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 208–216, 2018.

[24] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Oct. 2016, pp. 262–270.

[25] R. Samanta, R. Misir, and M. Mitra, "A reduced set of features for chronic kidney disease prediction," *J. Pathol. Informat.*, vol. 8, no. 1, p. 24, 2017.

[26] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *J. Biomed. Informat.*, vol. 53, pp. 220–228, Feb. 2015.

[27] L. J. Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 3, pp. 660–673, Sep. 2020.

[28] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *IAENG Transactions on Engineering Technologies*. Dordrecht, The Netherlands: Springer, 2013, pp. 245–257.

[29] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, nos. 7–9, pp. 1483–1493, Mar. 2009.

[30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[31] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[33] D. W. Hosner and S. Lemeshow, *Applied Logistic Regression*. New York, NY, USA: Wiley, 1989.

[34] V. Vapnik, *Statistical Learning Theory*, vol. 1, no. 624. New York, NY, USA: Wiley, 1998, p. 2.

[35] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge, 2000.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[37] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[39] L. S. Shapley, "A value for N-person games," *Contrib. Theory Games*, vol. 2, no. 28, pp. 307–317, 1953.

[40] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," 2017, *arXiv:1706.06060*. [Online]. Available: http://arxiv.org/abs/1706.06060

[41] T. Lancet Respiratory Medicine, "Opening the black box of machine learning," *Lancet Respiratory Med.*, vol. 6, no. 11, p. 801, Nov. 2018.

[42] C. Brugnara and K.-U. Eckardt, "Hematologic aspects of kidney disease," in *Brenner and Rector's the Kidney*, 9th ed. Philadelphia, PA, USA: Saunders, 2011, pp. 2081–2120.

[43] J. J. McMurray *et al.*, "Kidney disease: Improving global outcomes (KDIGO) anemia work group. KDIGO clinical practice guideline for anemia in chronic kidney disease," *Kidney Int. Suppl.*, vol. 2, no. 4, pp. 279–335, 2012.

[44] S. A. Olasunkanmi *et al.*, "Haematological profile of patients with chronic kidney disease in nigeria," *J. Nephrol. Renal Transplantation*, vol. 5, no. 1, pp. 2–10, 2012.

[45] S. B. Ghaderian and S. S. Beladi-Mousavi, "The role of diabetes mellitus and hypertension in chronic kidney disease," *J. Renal Injury Prevention*, vol. 3, no. 4, p. 109, 2014.

[46] E. J. Lamb, A. S. Levey, and P. E. Stevens, "The kidney disease improving global outcomes (KDIGO) guideline update for chronic kidney disease: Evolution not revolution," *Clin. Chem.*, vol. 59, no. 3, pp. 462–465, Mar. 2013.

[47] A. Forbes and H. Gallagher, "Chronic kidney disease in adults: Assessment and management," *Clin. Med.*, vol. 20, no. 2, p. 128, 2020.

[48] L. A. Inker *et al.*, "KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD," *Amer. J. Kidney Diseases*, vol. 63, no. 5, pp. 713–735, May 2014.

[49] M. Kinaan, H. Yau, S. Q. Martinez, and P. Kar, "Concepts in diabetic nephropathy: From pathophysiology to treatment," *J. Renal Hepatic Disorders*, vol. 1, no. 2, pp. 10–24, Jun. 2017.