This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

XGBoost Model for Chronic Kidney Disease Diagnosis

Adeola Ogunleye and Qing-Guo Wang

Abstract—Chronic Kidney Disease (CKD) is a menace that is affecting 10% of the world population and 15% of the South African population. The early and cheap diagnosis of this disease with accuracy and reliability will save 20,000 lives in South Africa per year. Scientists are developing smart solutions with Artificial Intelligence (AI). In this paper, several typical and recent AI algorithms are studied in the context of CKD and the extreme gradient boosting (XGBoost) is chosen as our base model for its high performance. Then, the model is optimized and the optimal full model trained on all the features achieves a testing accuracy, sensitivity and specificity of 1.000, 1.000 and 1.000, respectively. Note that, to cover the widest range of people, the time and monetary costs of CKD diagnosis have to be minimized with fewest patient tests. Thus the reduced model using fewer features is desirable while it should still maintain high performance. To this end, the set-theory based rule is presented which combines a few feature selection methods with their collective strengths. The reduced model using about a half of the original full features performs better than the models based on individual feature selection methods and achieves accuracy, sensitivity and specificity of 1.000, 1.000 and 1.000, respectively.

Index Terms—Medical diagnosis, Chronic Kidney Disease, Artificial Intelligence, Extreme Gradient Boosting, Clinical Decision Support System.

1 Introduction

Due to shortage and high pay of the specialists for manual disease diagnosis, a cheap and fast Clinical Decision Support System (CDSS) for automatic diagnosis is very beneficial to combat diseases. The Artificial Intelligence (AI) Algorithms such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), Naive Bayes, decision tree, the extreme gradient boosting one (XGBoost), logistic regression and fuzzy set theory incorporated with physician experience go a long way. They have been applied to diagnose diseases such as lung cancer, tuberculosis, cardiovascular diseases and malaria [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. The systems designed to accomplish such a diagnosis task are classified into the rule-based and the non-rule-based ones. An expert rule-based system is a system in which an expert system designer consults a number of experienced domain experts of interest to acquire knowledge about the task at hand before modelling it. This system contains an inference engine, a User Interface (UI) or shell to interact with the system and a knowledge-base to store the linguistics and rules of the system. Popular expert systems include DoctorMoon for tuberculosis and lung cancer, computer-assisted medical diagnosis (CADIAG) for heart diseases, MYCIN for bacterial infections and TxDENT 2.0 for dental diseases. Some of these expert systems have performed exceptionally well to the extent that they can diagnose several diseases, example of such is DoctorMoon which could diagnose ailments such as pulmonary tuberculosis, lung cancer, asthma, pneumonia and bronchiectasis con-currently [16]. The non-rule-based system involves no rules from a domain expert. A typical example of this type is ANN which infers its knowledge from continuous training on a data set. ANN has been used for disease diagnosis in [3], [6], [17], [18], [19]. A number of intelligent systems combine ANN with other ML techniques such as fuzzy logic system [20], [21]. The system performance has to be optimal to prevent over-fitting or under-fitting, where various optimisation techniques such as Genetic Algorithm (GA), Simulated Annealing (SA) and Particle Swarm Optimisation (PSO) are employed to tune the characteristics of the system towards the desired goal.

One focus area of this trend in the AI based medical diagnosis is Chronic Kidney Disease (CKD). CKD is the gradual loss in kidney's ability to filter the blood stream and rid off metabolic waste. Over 10% of the world's population and 15% of South Africa's population are affected by this disease [22]. Annually, kidney related ailments cost about one trillion dollars, thereby causing black market of kidney sales for patients in need of kidney transplant [23]. The main causes of CKD are over-weight, cardiovascular disease, high blood pressure, hereditary and delayed use of anti-retroviral drugs in HIV patients in world at large [24]. Chinese medicine, traditional herbs, alcohol, crystal-meth, consumption of other harmful substances and unhealthy eating habits can severely impair the kidney. The ailing patients condition go unnoticed because of the overlapping symptoms of the kidney disease with other ones. Fatigue and swollen feet in patients are the initially noticed symptoms of CKD. Diabetic patients are highly likely to have CKD, and CKD can be grouped into five different stages. A CKD patient in acute stages responds to treatment and can be revived. These acute stages can be regarded as Non-Dialysis-Dependent Chronic Kidney Disease (NDD-CKD) which extends from stage one to stage four. The last stage is stage five and known as End-Stage

- *Qing-Guo WANG, Institute for Intelligent Systems, Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa, E-mail: wangq@uj.ac.za*
- *Adeola OGUNLEYE, Institute for Intelligent Systems, Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa, E-mail: ogunleyeadeola7@gmail.com*

. Qing-Guo WANG and Adeola OGUNLEYE acknowledge the financial support of the National Research Foundation of South Africa (Grant Number: 113340) and Oppenheimer Memorial Trust grant, which partially funded their research on this work.

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Kidney Disease (ESKD). The patients at this stage require dialysis and intensive care. Whites and Indian population of South Africa and of the world do not suffer from CKD as much as the black population of African [25]. There is an urgent need for a very reliable Clinical Decision Support System (CDSS) to effectively diagnose CKD in South Africa because of the amount of yearly patients recorded [22]. The financial cost of dialysis for an end-stage CKD patient in South Africa amounts to about R120,000 ($9200) per year [26]. Early diagnosis of the disease can drastically prevent end-stage CKD, there-by reducing the number of patients on dialysis. Powerful and effective models that can diagnose CKD to assist Nephrologists and other medical practitioner is imperative. To develop a AI solution for CKD, one needs data collection and mining. Note that there is an overlap in the symptoms of CKD with other diseases and there is a need to narrow down the symptoms to the most important ones (features) so that patients will not need to carry out numerous tests to diagnose the presence of CKD. Thus, feature selection and reduction becomes important to choose most important ones for data mining.

On the other hand, significant scholarly works have been carried out on data mining part of the AI system on CKD, including but not limited to the following.

- Al-Hyari *et al.* [27] created models with Decision Tree (DT), Artificial Neural Network (ANN) and Naive Bayes(NB). The model accuracies for DT, NB and ANN are 92.2%, 88.2% and 82.4%, respectively. One hundred and two data points were used but could be too small with over-fitting.
- Deng *et al.* [28] identified the stages of Kidney Renal Cell Carcinoma (KIRC) with gene expression combined with DNA methylation data generating a fused network. A patient's network was first constructed from each type of data, followed by a fused network based on network fusion methods. Their model had an accuracy of 0.852.
- Salekin and Stankovic [29] designed the models using K-NN, Random Forest and Neural Network. They replaced the missing data in the dataset with values from *IBK* algorithm for the KNN model and the Neural Network. The features for their model were further reduced to create a minimised model for fast computation.
- Zheng *et al.* [30] took the transfer-learning method to extract imaging features from ultra sound kidney images in order to improve the diagnosis of congenital abnormalities of the kidney and urinary tract (CAKUT) diagnosis in children. Support vector machine was used to classify the features such as a combination of transfer learning features and conventional imaging features. The accuracy and area under ROC of the model is 0.87 and 0.92, respectively.
- Gupta *et al.* [31] determined the correlation between eleven chronic diseases including kidney disease. Several algorithms were used to diagnose Chronic Kidney Disease and AdaBoost algorithm performed with the best accuracy of 88.66%.

In this paper, we address feature reduction, learning algorithm selection and tuning for CKD. Our contributions are highlighted as follows:

- Unlike the domain models such as ANN and ADABoost, several ML techniques are compared on the CKD dataset without hyperparameter optimisation, and the XGBoost model is found to perform best. The XGBoost model is then used as our base model. The XGBoost model is then optimised, its performance based on the full statistical metrics is analysed and far better than the existing models.
- Some set-theoretical rule is presented to combine the positive expertise of a few feature selection techniques to obtain the final features used to train the model. This can be compared to the common clinic practice of employing a few Nephrologists to diagnose CKD and take a collective decision which should be better than individual ones in general. This rule works just so as expected.
- A minimised XGBoost model is then developed with the reduced feature set obtained by the above rule. Note that this reduced set has about a half of features as in the full set. Yet, the model trained on it yields the same performance as that of the full model with perfect prediction. Thus, this minimised CKD XGBoost model exhibits potency, and it can be embedded in a Clinical Decision Support System (CDSS) due to its fast, efficient and a low cost CKD diagnosis for patients in South Africa and other countries.

This paper is organised as follows. Section II describes the model development while Section III shows the model testing and benchmarking against other methods. Section IV concludes the paper and it emphasizes the salient points.

2 The Model development

The dataset is represented as D={ (x_i,y_i), i = 1, 2, ..., N}, where

$$x_i = \left[x_{i1}, x_{i2}, x_{i3}, ..., x_{ip}\right] \tag{1}$$

is a row vector with input variables (or features) of real-value as its elements, and

$$y_i \in \{0, 1\} \tag{2}$$

is a scalar with the output of integer-value as its element. The task in hand is a binary classification problem, that is, generate a model, $y = f(x)$, based on the training data. Then we can apply the model on the test data to predict $\hat{y_k}$=f(x_k) in hope that the predicted output $\hat{y}_k$ is same as the true output y_k for as many test points as possible.

The proposed model is shown in Figure 1. The first issue in model building via Machine Learning (ML) is feature selection/reduction. It is costly to collect more features and cumbersome to use them in model training. It is then imperative to reduce the features by selecting only the relevant and important features contributing to good output prediction. This also helps ML algorithms

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

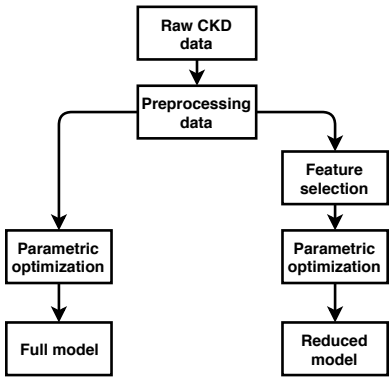

Fig. 1: Proposed method.

to learn hidden patterns in a dataset, reducing noise and correlations of some features. There exist many methods on feature selection. Each has their own strength and weakness. No one seems universally best. We searched for suitable methods according to our particular application of CKD and run screening simulation. Finally, the following three methods were chosen.

Recursive Feature Elimination (RFE). RFE recursively reduces the features in a dataset with repeated modelling [32]. RFE algorithm initially uses all the features to fit a model. The features will in turn be ranked according to the order of their importance [33]. Let L be a sequence of ordered numbers representing the number of features to be kept (L1 $>$ L2 $>$ L3 ...). Estimators are trained and each feature is assigned a particular weigh, after each round of recursive model creation, the features with the smallest weights are removed. Features that performed very well are kept. RFE is a meta-algorithm that uses other ML algorithm for its feature selection. Our RFE is based on the logistic regression.

Extra Tree Classifier (ETC). It is also known as Extreme Randomised Classifier (ERC) and was proposed by Geurts *et al.* [34]. ETC builds an ensemble of unpruned decision or regression trees in a top-down manner. The procedure is performed based on the feature representation and the partition to the right and left nodes. A tree grows until a specified tree depth. During the bagging process and with each attribute split, a random subset of features is used. Its two main differences from other tree-based ensemble methods are that it splits node by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees. Tree based classifiers have in-built feature selection. ETC can be used to determine the important features in the dataset. The more a particular feature is involved in the splitting process in the ETC model, the more import the feature is. This algorithm uses Gini index as its split criteria and uses greedy optimisation to perform its task.

Univariate Selection (US). Univariate Selection works by selecting features that have performed well in accordance to univariate statistical tests such as Pearson's Correlation, LDA, ANOVA and Chi-Square [35]. Chi-Square can be defined as a statistical test to evaluate the likelihood of correlation or association between features using their frequency distribution. For US technique, we invoked "SelectKBest" from SciKit-learn Python package [33], set scoring function and K to Chi2 and Chi4, respectively.

We rank all the features using RFE, ETC and US, respectively, to find good candidate features. Note that it is quite risky to base the final feature selection on any single method due to its weakness. Combining two or more methods which complement each other can effectively reduce weakness of individual methods. Thus our final selection rule is that feature j is taken if it is selected by two or three feature selection methods mentioned above.

The original (or full) dataset and the reduced dataset using the above feature selection procedure will be compared in the next section on the CKD case. It is also worth mentioning that our feature selection procedure can have many easy extensions to suit wide range of applications. For example, one can have 4-3 rule, that is, use four feature selection methods and take a feature which is selected by three or more methods. The key principles which should be always followed are that two or more methods should be used for preliminary selection and these methods should complement each other in sense that they do not share common weakness.

Once the suitable features have been obtained, we have to choose a ML algorithm to train a model function, $y = f(x)$. There has been great progress on ML over the decades. The new trend is deep learning which however usually needs a dataset containing several thousands of instances. For medical cases like the CKD data for a particular region, the data size is much smaller, especially after taking away invalid data points. It is known that tree-based algorithms, SVM and regression performs with smaller dataset. Thus, we look at the standard ML algorithms as well their new advances. We run a preliminary test for logistic regression, Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and extreme gradient boosting algorithm (XGBoost) on the CKD dataset and evaluated their performances. XGBoost was appealing because its default setting had a better average performance compared with the other classifiers in our test. We then chose it and optimise its performance.

XGBoost: Chen *et al.* [36] developed the XGBoost algorithm. It is an extendible and cutting-edge application of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms. It was developed for the sole purpose of model performance and computational speed. Boosting is an ensemble technique in which new models are added to adjust the errors made by existing models. Models are added recursively till no noticeable improvements can be detected. Gradient boosting is an algorithm in which new models are created that predict the residuals of prior models and then added together to make the final prediction. It uses a gradient descent algorithm to minimise the loss when adding

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

new models. This approach supports both regression and classification. XGBoost successful won 17 out of the 29 ML tasks posted on Kaggle by 2015. Using multiple core of a CPU and reducing the look up times of individual trees created in XGBoost, the performance was significantly improved. This algorithm is written in R and Python SciKit-learn [33] library and comes with new regularisation techniques.

To achieve optimal performance, the model has to be tuned carefully. Tuning XGBoost can be a very daunting task because of the number of hyperparameters it has. These parameters can be grouped into general, booster, learning task and command line parameters. Tuning can be done in a grid or random search. This paper uses the grid search. Grid search for optimum with a high dimension of parameters could be difficult. This can be easily managed by taking a smaller combination of parameters with reasonable ranges of parametric values at a time. K-fold cross validation is adopted to evaluate the model performance during model selection phase. For our simulation, Python packages and libraries were used because it is free and open source. The grid search is carried out as follow.

- Step 1: Keep 10% of the dataset as the validation set and another 10% as the test set. Use the remaining 80% to form our optimised XGBoost CKD model. The "n_estimators" which determines the epoch of the model is set to 100 and early_stopping_rounds to 10 to check for over-fitting.
- Step 2: Search for the optimal learning_rate and gamma simultaneously because they directly affect the performance of the model. The grid values searched for the learning_rate are 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, and 0.3, while those for the gamma are 0.1, 0.2, 0.5, 1, 1.5, 2, and 10. All the possible combinations of these two parameter values are run for the model tuning and the one with best performance is retained as the optimal values.
- Step 3: With the optimal values of the learning_rate and gamma, make a grid-search over the max_depth and min_child_weight in selected ranges of 1 to 10.
- Step 4: Make a grid-search over the L2 regularisation parameter reg_lambda and subsample simultaneously in selected ranges 0.1 to 1.
- Step 5: Make a grid-search over max_delta_step to adjust for imbalance in the dataset, grid searching value of 1 to 5.
- Step 6: Re-examine the model by simultaneous grid search over gamma, reg_lambda and subsample to check for differences between the optimum values.

Note that K-fold cross validation is used to avoid over-fitting. We take the most popular one of $K = 10$. The data is divided equally to 10 folds, 9 folds are used for training and the remaining one fold is for evaluation. The data set is reshuffled and the cross validation is repeated. After the training, the model prediction is assessed on the test data only. Thus the model performance such as the prediction accuracy and standard deviation can be obtained for comparison of the different ML algorithms and different parameter settings.

Note also that the prediction accuracy cannot be the only yardstick to select a classifier. Other criteria should also be considered. We follow the standard definitions of all the performance measures. Here are the basic terms:

- True Positive (TP): This is defined as the number of ill cases that are correctly predicted as ill.
- False Positive (FP): This is defined as the number of healthy cases that are wrongly predicted as ill.
- True Negative (TN): This is defined as the number of healthy cases that are correctly predicted as healthy.
- False Negative (FN): This is defined as the number of ill cases that are wrongly predicted as healthy.

Confusion matrix is a tabular representation of the above four numbers and exhibited in Table 1. The above terms are utilised to form several performance measures:

- Accuracy in ML is a metrics that is used to determine the amount of a particular class that is correctly predicted over the total number of sample. It can be calculated as $\frac{TP+TN}{TP+TN+FT+FN}$.
- Precision is the ratio of the number of correctly predicted ill cases over the total number of the predicted ill cases and calculated as $\frac{TP}{TP+FP}$
- Sensitivity is the ratio of the number of correctly predicted ill cases over the total number of the ill cases and calculated as $\frac{TP}{TP+FN}$.
- Specificity is the ratio of the number of correctly predicted healthy cases over the total number of the healthy cases and calculated as $\frac{TN}{TN+FP}$.
- Receiver Operating Characteristic (ROC) curve is the graph of True Positive Rate (TPR) against False Positive Rate (FPR). It shows the diagnostic ability of a model.
- Area Under Curve (AUC) is the area under ROC curve. AUC gives the sum measured performance across all possible classification thresholds.

TABLE 1: Confusion matrix

		Predicted class	
		Ill	**Healthy**
Actual class	**Ill**	TP	FN
	Healthy	FP	TN

3 THE MODEL SIMULATION

The online repository of University of California Irvine (UCI) provides a CKD dataset. This set contains 400 patients with 250 CKD cases and 150 CKD-free cases. By the ratio of the CKD-free and CKD patients in the dataset, the dataset is fairly balanced. The features are exhibited in Table 2. The patient's age (Age) lies between $2 - 90$ years old. The blood pressure (BP) lies between $50 - 120$ mmHg, while $70 - 490$ mgs/dl is the range of the blood glucose random (BGR). The blood urea (BU) lies between $15 - 424$ mgs/dl. The patients Serum Creatinine (SC) lies between $0.5 - 48.1$ mgs/dl, Sodium (SOD) in $4.5 - 150$ mEq/L and Potassium (POT) in $2.5 - 7.6$ mEq/L. Patient's haemoglobin (HEMO)

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

lies between $5.6 - 17.7$ gms, and their Pack Cell Volume (PCV) ranges in $16 - 53$, white blood cell count (WBCC) in $2200 - 26400$ cell/cumm, with their red blood cell count (RBCC) in $2.1 - 8.0$ millions/cumm. Specific Gravity (SG) is between $1.005 - 1.025$ while albumin and blood sugar lies between $0 - 5$. Four rows contain outliers and erroneous data, and these rows are removed in preparation for modelling [37], [38]. Then the number of data points/rows becomes 396. The categorical features in the dataset were converted to binary variables, that is, converting 'poor' or 'good', 'no' or 'yes', 'notpresent' or 'present' to 'zero' and 'one'. Some columns that are not binary were normalised to enable an ML algorithm to work consistently on it, although, some ML algorithms such as trees are immuned to normalisation. One uses

$$z_{ij} = \frac{x_{ij} - min(x_{.j})}{max(x_{.j}) - min(x_{.j})} \tag{3}$$

so that the normalised data in terms of z has the element value between 0 and 1. We tried dropping the rows with missing data making the instances drop from 396 to 157, which is a considerable reduction in the dataset and would badly affect the model due to a small data size. We considered several options of replacing missing data such as fill-forward and fill-backward, mean, median and mode imputation. The Gaussian curve was plotted to view the dataset after imputation, leading us to choose the median imputation because it formed a regular bell curve. In the original dataset, the instances of the same class are grouped together. We decided to shuffle the dataset to break the definite pattern. The resulting set is called the full dataset.

TABLE 2: CKD attributes or features

Number	Attribute	Abbreviation
1	Age	Age
2	Appetite	APPET
3	Anemia	ANE
4	Albumin	AL
5	Bacteria	BA
6	Blood Pressure	BP
7	Blood Glucose Random	BGR
8	Blood Urea	BU
9	Coronary Artery Disease	CAD
10	Diabetes Mellitus	DM
11	Pus Cell Clump	PCC
12	Serum Creatinine	SC
13	Sodium	SOD
14	Potassium	POT
15	Hemoglobin	HEMO
16	Pack Cell Volume	PCV
17	White Blood Cell Count	WC
18	Red Blood Cell Count	RBCC
19	Hypertension	HTN
20	Pus Cell	PC
21	Specific Gravity	SG
22	Red Blood Cell	RBC
23	Petal Edema	PE
24	Sugar	SU
25	Class	CLASS

There is a large number of ML algorithms. The size of the dataset makes complex techniques such as ANN unsuitable because of over-fitting. The state-of-the-art use of ANN is deep learning and it performs well with a large dataset. It is known that tree-based algorithm, SVM and regression perform well with small dataset, among them, we chose the methods of K-Nearest Neighbour (KNN), logistic regression (LG), Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Support Vector Machine (SVM) and XGBoost (XGB). They were applied to the full dataset, respectively, with the normal/default parameters without tuning. Table 3 shows the performance metrics including accuracy, Root Mean Square Error (RMSE), f1_score, precision, sensitivity and Area Under Curve (AUC). Overall, XGBoost has the best performance and thus it was taken forward as our ML algorithm.

TABLE 3: Performance of AI methods.

Abbreviation	Accuracy	F1_score	Sensitivity	AUC	RMSE
KNN	0.968±0.032	0.93±0.10	0.87±0.16	0.98±0.04	0.18±0.18
LG	0.981±0.026	0.95±0.10	0.91±0.16	1.00±0.00	0.14±0.17
SVM	0.981±0.026	0.95±0.10	0.91±0.16	1.00±0.00	0.14±0.17
LDA	0.981±0.025	0.95±0.10	0.93±0.16	1.00±0.00	0.14±0.20
CART	0.981±0.016	0.96±0.07	0.95±0.10	0.99±0.04	0.14±0.17
XGBoost	0.987±0.016	0.97±0.06	0.98±0.08	1.00±0.00	0.11±0.16

We next refined the XGBoost model on the full features with the parameter optimisation method of Section 2. The model was trained and tuned on the training dataset. To avoid over-fitting of the model, an early-stop parameter was set to 10. The classification error and loss function of the optimisation process is shown in Figures 2 and 3. The optimal values of hyperparameters were obtained and are given in Table 4. The validation dataset was used to check the performance of the trained model. If it is satisfactory, we proceed to test. The test set was used to evaluate the model. Performance of this model is compared in Table 5 with the CKD models found in the literature. Based on these measures, the proposed model gives significant improvement.

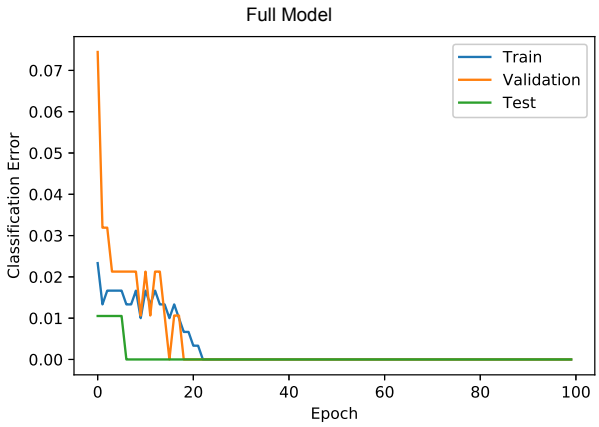

Fig. 2: Classification error of the full model.

Correlated features are redundant and may degrade the performance of ML algorithms. The correlations between the features are depicted in Figure 4, which is from [39]. Thus, we had to make the feature reduction. The twenty-four features were ranked using the Recursive Feature Elimination, Extra Tree Classifier and Uni-variate Selection,

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

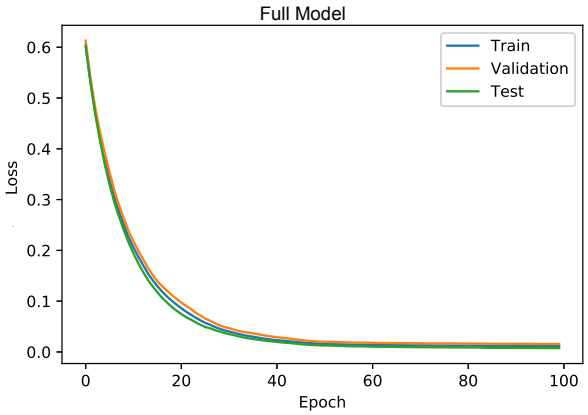

Fig. 3: Loss function of the full model.

TABLE 4: Optimal parameter values for full model.

Hyperparameter	Value
learning_rate	0.1
n_estimator	100
gamma	0.1
reg_lambda	0.2
subsample	0.8
min_child_weight	1
max_depth	4
max_delta_step	2

TABLE 5: Comparison of models.

Model	Accuracy	Sensitivity	Specificity
Gupta *et al.*	0.886	Nil	Nil
Al-Hyari *et al.*	0.922	0.939	0.842
Park *et al.*	0.831	0.900	0.696
Full model	1.000	1.000	1.000

respectively. The ranking is shown, respectively, in Figures 5 - 7, where one sees that the features are ranked differently by different algorithms. A cut-off point is determined to select the top-ranked features for the best trade-off between model performance and simplicity. The cut-off points for RFE, ETC and US are 10, 2.5 and 2.5, respectively. After this selection, the sets of the retained features for RFE, ETC and US are represented by S_1, S_2 and S_3, respectively. S_1 has 15 features while S_2 and S_3 has 13 features each respectively.

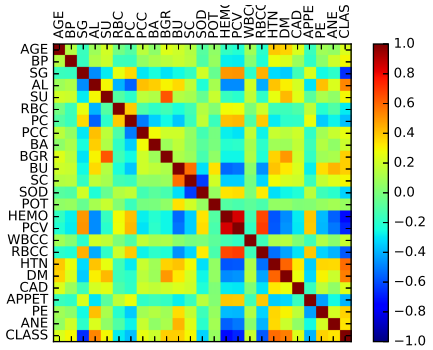

Fig. 4: Correlation plot for CKD attributes.

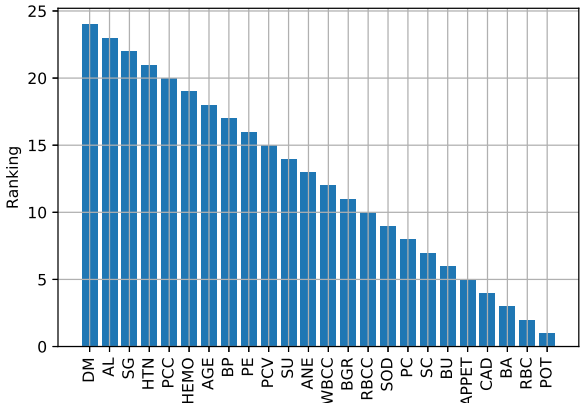

Fig. 5: Ranking of CKD features with RFE.

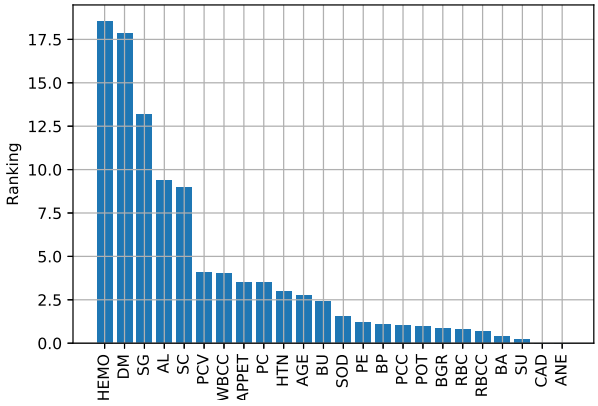

Fig. 6: Ranking of CKD features with ETC.

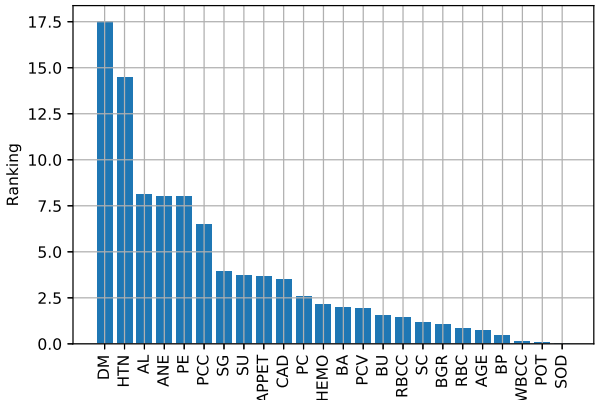

Fig. 7: Ranking of CKD features with US.

S_1 has 15 features:

$$S_1 = \{DM, AL, SG, HTN, PCC, HEMO, AGE, BP, PE, PCV, SU, ANE, WBCC, RBCC, BGR\}, \quad (4)$$

S_2 has 13 features:

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

$$S_2 = \{HEMO, DM, SG, AL, SC, PCV, APPET, WBCC, PC\}, \quad (5)$$

and S_3 has 13 features:

$$S_3 = \{DM, HTN, AL, ANE, PE, PCC, SG, SU, APPET, CAD, PC\}. \quad (6)$$

TABLE 6: Optimal parameter values for the selected features.

Hyperparameter	RFE	ETC	US	Reduced
Learning_rate	0.1	0.3	0.1	0.2
N_estimator	100	100	100	100
gamma	0.5	0.2	0.1	0.3
reg_lambda	0.2	0.6	0.3	0.2
subsample	0.5	0.5	0.8	0.7
min_child_weight	2	1	2	2
max_depth	3	3	2	3
max_delta_step	2	1	2	2

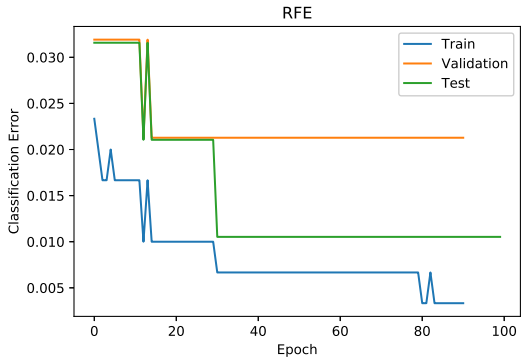

Fig. 8: Classification error of RFE model.

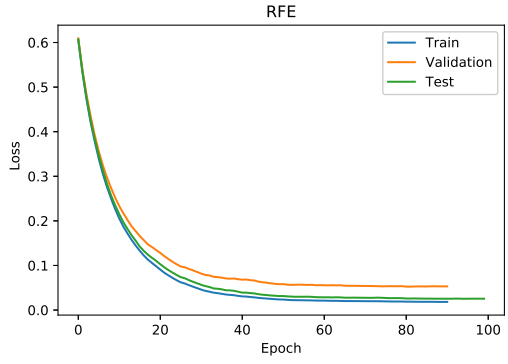

Fig. 9: Loss function of RFE model.

The optimal XGboost modelling was applied to each S_i, respectively. The tuning process is shown in Figures 8-13. The optimal values of the hyperparameters are given in Table 6. Three models work like three experienced Nephrologists to diagnose patients separately. Their performance is compared in Table 7. Out of the three feature selection algorithm, RFE performed best and it has an accuracy, precision and sensitivity of 0.989, 0.985 and

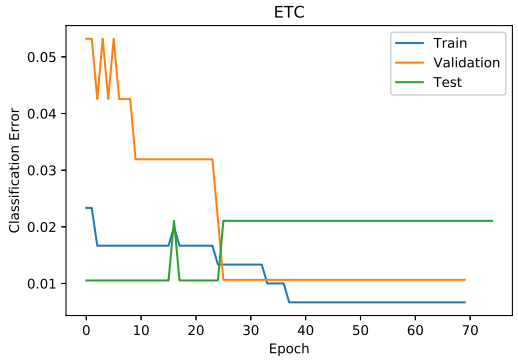

Fig. 10: Classification error of ETC model.

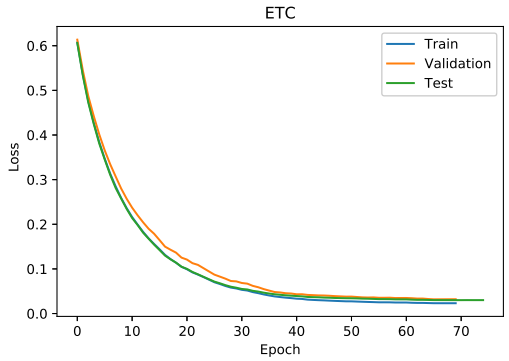

Fig. 11: Loss function of ETC model.

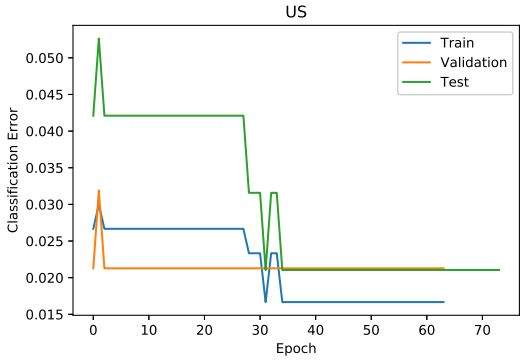

Fig. 12: Classification error of US model.

1.000 respectively, but worse than the full model which has perfect scores for all the measures.

Finally, we applied our feature selection rule in Section 2 to get our reduced set of the feature S_r: a feature is taken only if it is a member of at least two sets of the above three. This yields

$$S_r = \{DM, AL, SG, HTN, HEMO, WBCC, AGE, PCV, ANE, PE, SU, APPET, PC\}.$$

which has 13 features only. The optimal XGboosting modeling was applied to S_r. The tuning process is shown in Figures 14 and 15, while the optimal parameter values

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

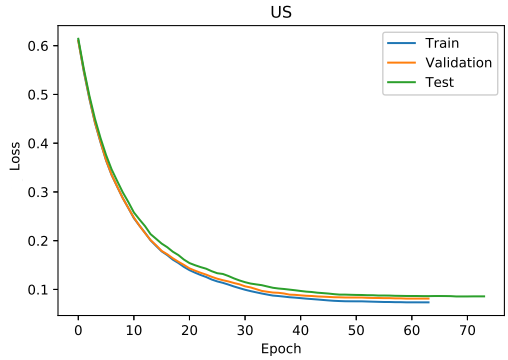

Fig. 13: Loss function of US model.

are given in Table 6. The resulting reduced model is then compared in Table 7 with the full model and the models obtained above with individual feature selection methods. It has perfect scores for all the measures. Note that if a smaller set of features is used effectively to build a model to diagnose CKD, it will reduce the money and time that the patient spends on medical tests. Besides, it will also make the diagnosis system simpler and faster.

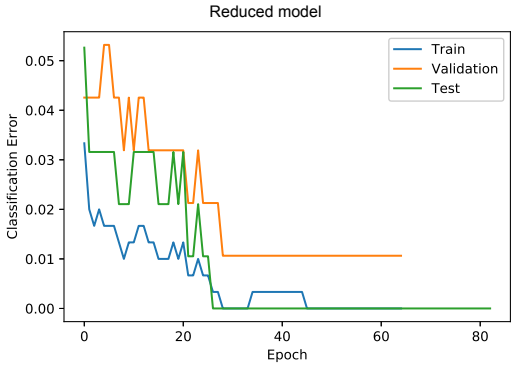

Fig. 14: Classification error of the reduced model.

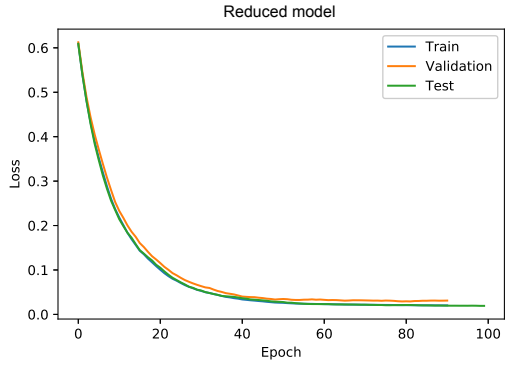

Fig. 15: Loss function of the reduced model.

To characterise our full model (based on full features) and reduced model (based on finally selected features), their performance was further studied with the ROC plot in Figures

TABLE 7: Model comparison.

Model	Accuracy	Precision	Sensitivity	Specificity	MAE
RFE	0.989	0.985	1.000	0.983	0.011
ETC	0.979	0.981	0.981	0.979	0.021
US	0.979	1.000	0.969	0.974	0.021
Reduced model	1.000	1.000	1.000	1.000	0.000
Full model	1.000	1.000	1.000	1.000	0.000

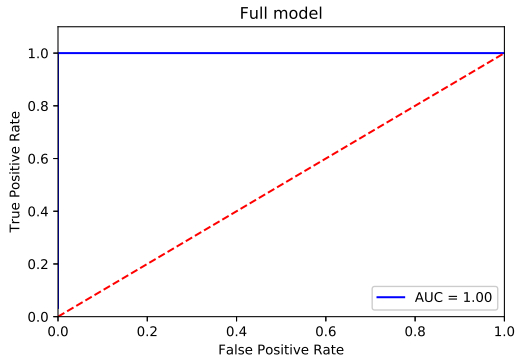

Fig. 16: ROC of the full model.

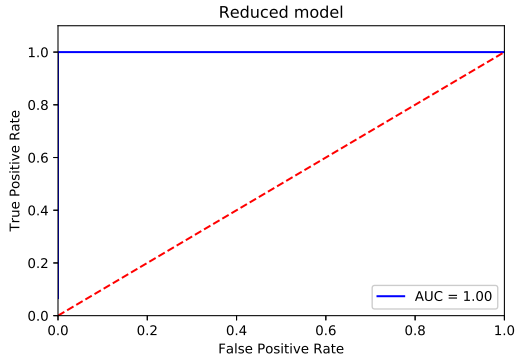

Fig. 17: ROC of the reduced model.

16 and 17, where two cases show invisible differences. An out-of-sample set of about 10% dataset was kept to test our model's predictive power. This test set translates into 40 patients, of which 11 patients had CKD while the rest are CKD-free. The full model correctly diagnosed all the 11 CKD and all the 29 CKD-free patients, which is perfect. The confusion matrix is shown in Table 8. On the other hand, the reduced model matched the full model performance-wise. Its confusion matrix is shown in Table 9. Note that the numbers in the confusion matrix had been used to calculate the performance measures shown in Table 7.

TABLE 8: Confusion matrix for the full model.

		Predicted class	
		CKD	NoCKD
Actual class	**CKD**	11	0
	NoCKD	0	29

4 Conclusion

In this paper, the XGBoost method has been studied and optimised for CKD diagnosis. The resulting CKD models are

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 9: Confusion matrix for the reduced model.

		Predicted class	
		CKD	**NoCKD**
Actual class	**CKD**	11	0
	NoCKD	0	29

compared with the existing CKD models in the domain. The proposed full model has achieved an accuracy, sensitivity and specificity of 1.000, 1.000 and 1.000, respectively. Three feature selecting techniques are combined by leveraging the strengths of each technique. A reduced model with about a half of the full features has an accuracy, sensitivity and specificity of 1.000, 1.000 and 1.000, respectively. These models coupled with the experience of a Nephrologist can help reduce the cost and time to diagnose a CKD patient. It is noted that the some tests for CKD could produce images as raw data. The image based feature extraction and learning show great potentials in ML in general and medical application in particular [40]. It would be in future research to use it for CKD case.

REFERENCES

[1] C.-S. Lee and M.-H. Wang, "A fuzzy expert system for diabetes decision support application." *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 41, no. 1, pp. 139–153, 2011.

[2] C. B. Delahunt, C. Mehanian, L. Hu, S. K. McGuire, C. R. Champlin, M. P. Horning, B. K. Wilson, and C. M. Thompon, "Automated microscopy and machine learning for expert-level malaria field diagnosis," *Proceedings of the 5th IEEE Global Humanitarian Technology Conference, GHTC 2015*, pp. 393–399, 2015.

[3] B. D. Sekar, C. M. Dong, J. Shi, and X. Y. Hu, "Fused hierarchical neural networks for cardiovascular disease diagnosis," *IEEE Sensors Journal*, vol. 12, no. 3, pp. 644–650, 2012.

[4] S. Basnet and N. Venkatraman, "A novel fuzzy-logic controller for an artificial heart," *Proceedings of the IEEE International Conference on Control Applications*, pp. 1586–1591, 2009.

[5] C. Arya and R. Tiwari, "Expert system for breast cancer diagnosis: A survey," *2016 International Conference on Computer Communication and Informatics, ICCCI 2016*, pp. 1–9, 2016.

[6] J. Chen, X. Qi, O. Tervonen, O. Silvén, G. Zhao, and M. Pietikäinen, "Thorax disease diagnosis using deep convolutional neural network," *Cadence Whitepaper*, pp. 2287–2290, 2016.

[7] H. Sun, L. Zhang, X. Hu, and L. Tian, "Experiment study of fuzzy impedance control on horizontal lower limbs rehabilitation robot," *2011 International Conference on Electronics, Communications and Control (Icecc)*, pp. 2640–2643, 2011.

[8] Samuel and Omisore, "Hybrid intelligent system for the diagnosis of typhoid fever," *Journal of Computer Engineering & Information Technology*, pp. 1–5, 2013.

[9] W. Suparta and K. Alhasa, "Modeling of tropospheric delays Using ANFIS," in *Adaptive Neuro-Fuzzy Interference System*, 1st ed. Springer, 2016, no. 2009, ch. 2, pp. 5–18.

[10] N. Swain, "A Survey of application of fuzzy logic in intelligent transportation systems (ITS) and rural ITS," *Proceedings of the IEEE SoutheastCon 2006*, pp. 85–90, 2006. [Online]. Available: https://ieeexplore.ieee.org/document/1629329

[11] M. G. Tsipouras, C. Voglis, and D. I. Fotiadis, "A framework for fuzzy expert system creation–application to cardiovascular diseases." *IEEE transactions on bio-medical engineering*, vol. 54, no. 11, pp. 2089–2105, 2007.

[12] M. J. Er and S. Mandal, "A survey of adaptive fuzzy controllers: Nonlinearities and classifications," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 5, pp. 1095–1107, 2016.

[13] A. Bhatia, V. Mago, and R. Singh, "Use of Soft Computing Techniques in Medical Decision Making: A Survey," *International Conference on Advances in Computing,Communications and Informatics (ICACCI)*, pp. 1131–1137, 2014.

[14] D. Pacheco, "A Web-based Fuzzy Inference System Based Tool for Cardiovascular Disease Risk Assessment," *NOVA*, pp. 7–16, 2015.

[15] K. H. Park, K. S. Ryu, and K. H. O. Ryu, "Determining Minimum Feature Number Of Classification On Clear Cell Renal Cell Carcinoma Clinical Dataset," *International Conference on Machine Learning and Cybernetics*, pp. 894–898, 2016.

[16] N. H. Phuong and V. Kreinovich, "Fuzzy logic and its applications in medicine," *International Journal of Medical Informatics*, vol. 62, no. 2-3, pp. 165–173, 2001.

[17] M. Kumar, A. Sharma, and S. Agarwal, "Clinical Decision Support System for Diabetes Disease Diagnosis Using Optimized Neural Network," *Journal of Computational Science*, vol. 3, no. 5, pp. 254–261, 2014.

[18] E. O. Olaniyi, O. K. Oyedotun, and A. Helwan, "Neural Network Diagnosis of Heart Disease," *International Conference on Advances in Biomedical Engineering (ICABME) output*, pp. 21–24, 2015.

[19] R. F. Olanrewaju, N. S. Sahari, A. A. Musa, and N. Hakiem, "Application of neural networks in early detection and diagnosis of Parkinson's disease," *2014 International Conference on Cyber and IT Service Management (CITSM)*, pp. 78–82, 2014.

[20] A. Tzavaras, P. R. Weller, and B. Spyropoulos, "A neuro-fuzzy controller for the estimation of tidal volume and respiration frequency ventilator settings for COPD patients ventilated in control mode," *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pp. 3765–3768, 2007.

[21] K. Polat and S. Güneş, "Detection of ECG arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine," *Applied Mathematics and Computation*, vol. 186, no. 1, pp. 898–906, 2007.

[22] G. Pippa, "Chronic kidney disease on the rise in SA," *Life Healthcare*, pp. 1–3, 2015.

[23] C. Palace, L. Vegas, P. Kotanko, N. York, N. W. Levin, and C. Ronco, "Advances in chronic kidney disease," *International Journal of Molecular Sciences*, vol. 36, no. 3-4, pp. 147–150, 2016. [Online]. Available: http://www.karger.com?doi=10.1159/000357165

[24] M. R. Moosa, I. Van Der Walt, S. Naicker, and A. M. Meyers, "Important causes of chronic kidney disease in South Africa," *South African Medical Journal*, vol. 105, no. 4, pp. 1–8, 2015.

[25] S. Naicker, "End-stage renal disease in sub-saharan and South Africa," *Kidney International - Supplement*, vol. 63, pp. 119–122, 2003.

[26] K. Herrmannsen, "Thousands may die without life-saving dialysis," *HEALTH-E NEWS*, pp. 1–2, 2015. [Online]. Available: health-e.org.za /2015/03/12/thousands-may-die-without-life-saving-dialysis/

[27] A. Y. Al-Hyari, A. M. Al-Taee, and M. A. Al-Taee, "Clinical decision support system for diagnosis and management of Chronic Renal Failure," *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp. 1–6, 2013.

[28] S.-P. Deng, S. Cao, D.-S. Huang, and Y.-P. Wang, "Identifying Stages of Kidney Renal Cell Carcinoma by Combining Gene Expression and DNA Methylation Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 5, pp. 1147–1153, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7563822/

[29] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," *IEEE International Conference on Healthcare Informatics*, pp. 262–270, 2016.

[30] Q. Zheng, G. Tasian, and Y. Fan, "Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on Ultrasound imaging data," *International Symposium on Biomedical Imaging*, vol. abs/1801.0, no. Isbi, pp. 1487–1490, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7563822/

[31] D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 281–287, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7813730/

[32] J. Go and T. Łukaszuk, "Application of the Recursive Feature Elimination and the Relaxed Linear Separability Feature Selection Algorithms To Gene Expression Data Analysis," *Advances In Computer Science Research Application*, vol. 10, pp. 39–52, 2013. [Online]. Available: file:///Users/Littleyi/Downloads/Goscik.pdf

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot,

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics

and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 36, no. 1, pp. 3–42, 2006.

[35] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7160458/

[36] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016. [Online]. Available: https://arxiv.org/abs/1603.02754

[37] A. Birkett, "How to Deal with Outliers in Your Data," 2017. [Online]. Available: https://conversionxl.com/blog/outliers/

[38] N. Sharma, "Ways to Detect and Remove the Outliers," 2018. [Online]. Available: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

[39] O. A. Azeez and Q. G. Wang, "Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease," *I4th IEEE International Conference on Control and Automation, June 12-15, 2018*, pp. 805–810, 2018.

[40] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, and D. Zhang, "Multi-View Generative Adversarial Network and Its Application in Pearl Classification," *IEEE Transactions on Industrial Electronics*, vol. PP, no. c, pp. 1–9, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8575147/

Qing-Guo Wang received, respectively, B.Eng. in Chemical Engineering in 1982, M. Eng. in 1984 and Ph.D. in 1987 both in Industrial Automation, all from Zhejiang University, PR China. He held Alexander-von-Humboldt Research Fellowship of Germany from 1990 to 1992. From 1992 to 2015, he was with the Department of Electrical and Computer Engineering of the National University of Singapore, where he became a Full Professor in 2004. He is currently a Distinguished Professor with Institute for Intelligent Systems, University of Johannesburg, South Africa. He holds A-rating from the National Research Foundation of South Africa. He is a member of Academy of Science of South Africa. His present research interests are mainly in modelling, estimation, prediction, control, optimisation and automation for complex systems, including but not limited to, industrial and environmental processes, new energy devices, defense systems, medical engineering, and financial markets. He has published nearly 300 international journal papers and 7 research monographs. He received about 15000 citations with h-index of 65. He is currently the Deputy Editor-in-Chief of the ISA Transactions (USA).

Ogunleye Adeola received a bachelor of engineering degree from the Olabisi Onabanjo University in 2012. Later, he received his masters from Stellenbosch University in 2016. He is currently working towards his Ph.D. degree at the University of Johannesburg, sponsored by the National Research Foundation (NRF). His research interests include machine learning, computer vision, and the Internet Of Things. He is currently developing an automated diagnosis system for kidney disease using machine learning. He is also an embedded system designer and programmer with hands-on experience on complex project stages ranging from PCB designs to programming of chips and integration of systems.

1545-5963 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.