An Extreme Gradient Boosting-based Approach for Effective Chronic Kidney Disease Diagnosis

Bader Fahad Alkhamees[1*]

Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11362, Saudi Arabia

Abstract

Chronic kidney disease is one of the critical illnesses that affects roughly 10% of the people in the world. Early and accurate prediction of such disease is required for proper treatment. The use of machine learning (ML) for medical diagnosis in healthcare has increased. The doctor can identify the disease early with the aid of ML algorithms and approaches. This study aims to develop a diagnosis approach to recognize chronic kidney disease and assist the experts for exploring preventive measures early using extreme gradient boosting (XGBoost) model. The XGBoost is used due to its ability in-build features to manipulate missing data and its regularization capability to handle unbalanced datasets. The approach is trained and evaluated on a public dataset consisted of 24 features for 400 patients taken from the University of California Irvine (UCI) repository. The mean and most frequent values are used respectively for replacing the missing numerical and categorical values. The experimental results using a 10-fold cross-validation and holdout test techniques with a number of evaluation metrics exposed that the XGBoost model of the proposed approach achieves a competitive high result compared with the recent work on the same dataset. It attained 99.9% of AUC mean for the 10-fold cross-validation test and 99.6 of accuracy for 60% holdout test from the dataset to diagnosis the chronic kidney disease.

Keywords:

Machine Learning, Extreme Gradient Boosting (XGBoost) Model, Chronic Kidney Disease, Medical Diagnosis, Area Under-Curve (AUC).

1. Introduction

The term "chronic kidney disease" (CKD) describes a gradual decline in the construction and functions of people's kidneys, particularly the decline in filtration rates over months or years. The endocrine, excretory, and metabolic kidneys functions are gradually disappear as a result of abnormal biochemical changes that start off the process.

Renal failure symptoms and signs appear as a result of these abnormalities. The most frequent reasons of the disease are recorded as diabetes, hypertension, glomerular and interstitial diseases, inflammatory disorders, congenital conditions, and abnormalities of nonvascular [1]. However, the primary cause of the disease still remains unknown in many patients.

The amount of albumin in the urine and the glomerular filtration rate (GFR) are used to assess the disease's prognosis. Reduced GFR and higher levels of albumin in the urine have been linked to be a high risk leads to mortality, CVD mortality, advanced acute kidney injury and disease [2]. It was believed that atherosclerotic calcification in vessels tracked by the formation of cholesterol crystal was the cause of a patient's high risk for developing CKD [3].

If left untreated, chronic kidney disease (CKD) progresses over a pathological range of conditions to end stage renal failure (ESRF) or end stage renal diseases (ESRD) that can cause patients to go into a coma or pass away [4]. It can be challenging for the patients or the doctor for suspecting kidney involvement when CKD slowly develops because it either goes unnoticed or exhibits a variety of non-specific indications such as fatigue, weight loss, edema, poor appetite, headaches, and muscle cramps. In addition, some indications do not appear until greatly later, in the third stage or fourth stage of the kidney diseases, by which point comorbidity has already developed [5].

Hematopoietic abnormalities, immune dysfunction, endocrine dysfunction, electrolyte imbalance, and neurological symptoms are additional signs of CKD [1]. As a result, CKD raises the risk of developing other illnesses like the CVD mentioned above, increasing the number of deaths and morbidities [6]. Consequently, CKD has grown to be a major global burden and is responsible for a sizable portion of non-communicable disease-related fatalities (NCD). From 1990 to 2010, it rose from being the 27th to the 18th most common cause of death worldwide [7].

In 2013, about one million people passed away due to the CKD or conditions linked to it [8]. Over the past ten years, the total of new-cases who requiring renal-replacement therapy has been raised by a global rate equals 8% annually [9]. According to studies, CKD is more of a burden in low- and middle-income countries than in high-income ones [10], [11]. The authors in [12] indicated that CKD creates a high cost load to the systems of healthcare in the world. Moreover, the authors in [13] stated that the median prevalence of CKD is varied from 23.4% to 35.8% in aged 64 years or older.

Manuscript received September 5, 2022
Manuscript revised September 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.9.92

To ensure that a patient receives an effective course of treatment from the doctor, the situation necessitates the development of a diagnostic method, or, developing a screening system, for early and accurate CKD detection from data of patients. One of the most effective and notable methods in the medical sector to diagnose and predict several diseases and their stages is a machine learning (ML) methods [14-21]. The datasets of numerous diseases can be used to develop ML methods and models.

ML methods are widely used to analyse and explore the vast datasets and all of their patterns, features, modes, etc. [22-28]. Incorporating algorithms into medical databases will help professionals make educated decisions about illnesses, avoid mistakes, and ensure that the general public lives in safety [29].

Therefore, this paper aims to develop a diagnosis approach to recognize chronic kidney disease and assist the experts for exploring preventive measures early using extreme gradient boosting (XGBoost) model.

The reminder of the paper is organized as follows: Section 2 gives the literature review. The research methods is presented in section 3. Section 4 presents the results and discussion. Section 5 summarizes the conclusion and future work.

2. Literature Review

According to the previous studies, ML models have been utilized to perform satisfactorily results in this context. Many researchers and data scientists have employed a variety of techniques to detect kidney disease from the input data of patients. For example, K-star, SVM, and J48 algorithms were applied to a dataset taken from UCI by Engin et al., who then compared their sensitivity, accuracy, and other parameters. They found that the J48 model attained 99% of accuracy [30].

Contrarily, Gunarathne et al. [31] tested various procedures on the similar dataset and discovered that the Multi-class Decision Random Forest model outperformed them all with an accuracy rate of 99.1%. However, Nusrat et al. [32] used a different strategy when they featured datasets, pre-processing the information using some metrics such as the area under curve (AUC), root mean squared error, and mean absolute error. They applied the Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbour (KNN), and Support Vector Machine (SVM) algorithms after highlighting the dataset. They found that DT offers the best accuracy, ranging from 98% to 99%.

Moreover, Huseyin et al. [33] did a different study where, prior to using the algorithm, they made adjustments to the dataset's feature selection. The dataset was therefore subjected to the filter, embedded, and wrapper feature selection approaches before being subjected to the SVM algorithm. Their research indicates that the subset evaluation for the filter scheme achieved the best accuracy (98.5%). Devika et al. [34] introduced a study to predict CKD also concentrated on Nave Bayes (NB), Random Forest (RF) K-, and Nearest Neighbor (KNN) algorithms. With an accuracy of 99% among these classification models, RF outperformed the other models.

Additionally, Merve et al. [35] improved the accuracy using ensemble learning by the AdaBoost model. They used the root mean squared error (RMSE), mean absolute error (MAE), and area under the curve (AUC) in their work to evaluate the ML models' performance. Amanah et al. [36] also used the PSO algorithm for optimizing the results, and they raised the average of accuracy to 36.20% after combining the AdaBoost and PSO feature selection algorithms.

Alternatively, Chittora et al. [37] used some machine learning algorithms like linear support vector machine (LSVM), Chi-square automatic interaction detection (CHAID), artificial neural network (ANN), decision tree (C5.0), K-nearest neighbours (KNN), and random tree and six different feature selection methods, with a deep learning network achieving a high accuracy result. However, the authors did not give any detail about this accuracy result with other evaluation metrics like F1-score and did not include them in the table of comparison results with other ML models.

Furthermore, a study by Sobrinho et al. [38] examined how ML techniques can aid for early CKD detection in developing nations. The study's conclusions suggest that the J48 decision tree, with its 95.00% accuracy in classification, is a noble ML technique for this type of developing in screening countries.

Boosting algorithms were used in the study and attained a comparative result of accuracy with the AdaBoost model [39]. Therefore, it is clear that ML techniques open doors for early detection of CKD so that the patients can receive better care.

Almasoud and Ward introduced a chronic kidney disease detection approach using a number of machine learning models such as support vector machine (SVM), random forest (RF), Logistic regression (LR), and gradient boosting (GB) algorithms. They trained and tested these models using 10-fold cross-validation. They achieved an accuracy of 99% from GB model [40].

In order to improve the outcome of previous studies and methods, an effective and accurate approach will be developed to detect and recognize CKD early and assist the experts for exploring preventive measures and ensuring a proper treatment.

3. Research Methods

This section gives an explanation about the research methods used in the proposed approach. At first, the section describes the Classification and Regression Trees method. After that, it explains the Extreme Gradient Boosting (XGBoost) method.

3.1 Classification and Regression Trees

In 1984, Leo Breiman introduced the Classification and Regression Trees (CART) approach. It is a decision tree method that uses nonparametric statistics for its task [41]. The CART decision tree can be created by repeatedly splitting a node into its two-roots. The fundamental principle of the CART tree is to select the purest root node for the child produced by choosing between all possible splitting at each node. For instance, assuming training data T as shown below.

$$T = \{(x_i, y_i)\}, 1 \le i \le n$$
$$A = [x_{1,1\dots m} | \, x_{2,1\dots m} | \dots | x_{n,1\dots m}] \subset R^{nm}$$
$$x_i = [a_{ij} \in A], 1 \le j \le m$$
$$v_i = [a_{ij} \in A], 1 \le i \le n$$

Where n is the total number of cases of chronic kidney disease, m is the total number of predictors in a dataset, x_i is the value for each instance, v_i is the total number of values of the variables or predictors, and y_i is the actual target, or the outcome of the prediction of cases of chronic kidney disease.

For the multiclass classification, CART was constructed using training data in order to group training data that had the same level or value. In order to create the cleanest data partitions possible, the CART formation process involves choosing thresholds or splitting rules on features. As a result, the observations with same value or level are assembled on every branch or division of the resulting tree. For multi-level classification, resulted CART function is a binary-tree graph with a level, which links to the data input, in which, every node of the graph has an instance of data related to patients' chronic kidney disease. As a result, the function f in a dataset x can be formulated as follows:

$$f(x) = w_{q(x)}, q: R^m \to \{1, 2, \dots, T\}, w \in R^T$$
$$I_t = \{i | x_i \in t\}, 1 \le t \le T$$
$$w_t = \frac{|\{i \in I_t | y_i = 1\}|}{I_t}, 1 \le t \le T$$

Where T is the number of leaf nodes at f, $q(x)$ maps dataset instances to one of leaf nodes at f, I_t indexes the data that is on node t, and $w_t, 1 \le t \le T$ is the weight of leaf on f. A node's purity is determined using Gini impurity that is represented by:

$$G = 1 - \sum_{i=1}^{k} (p_i)^2 \quad (1)$$

Where $i = 1,2,3, \dots, k$, and k is a large number of levels with probability p_i. If the data on the node only has one level or if the data on the node has a similar or homogeneous level, the value of this root node will be minimum.

3.2 Extreme Gradient Boosting (XGBoost)

Bagging and boosting are the two main techniques used in the ensemble method. Prediction accuracy is increased sequentially at a given time using the bagging method, which involves building multiple models independently, in which the final output of prediction is the average of prediction outputs [42]. Creating several models sequentially while basing each model's error function on the performance of the one before it is known as "boosting." The accuracy of the weakest weak hypothesis's basic model, which is the smallest, determines how well the boosting method works [43].

The gradient boosting approach, which can be used for both classification and regression problems, is an illustration of a boosting strategy. Since the 1990s, the research on the gradient boosting algorithms has been conducted in a number of scientific fields. In 1984, Leo Breiman was the first to introduce gradient boosting, which can be seen as a suitable optimization algorithm [44]. The idea behind gradient boosting algorithms lies in their expansion, specifically the additional criterion fittings will be expanded, where the boosting process consists of sequentially minimizing Root Mean Square Error (RMSE) [45].

Extreme Gradient Boosting (XGBoost) is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning method. It involves creating a gradient boosting model with superior performance and processing speed. The benefit of processing missing value data with XGBoost is that no data is imputed at the start of learning [46]. By minimizing the objective function of regularization for each tree, the XGBoost method improves the performance of each tree created by the CART process as weak learners. Algorithm 1 gives split finding of exact greedy algorithm.

Algorithm 1: Exact Greedy Algorithm for Split Finding

Input: I: a set of instances in the current node, m: feature dimension

$g \leftarrow 0$ //gain

$G \leftarrow \sum_{i \in I} g_i$, and $H \leftarrow \sum_{i \in I} h_i$

$for\ k = 1\ to\ m$:

  $G_L \leftarrow 0, H_L \leftarrow 0$

  $for\ k\ in\ sorted\ do$:

    $G_L \leftarrow G_L + g_i,\ H_L \leftarrow H_L + h_i$

    $G_R \leftarrow G - G_L,\ H_R \leftarrow H - H_L$

    $score \leftarrow \max\left[score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]$

  end

end

Output: Split with max score

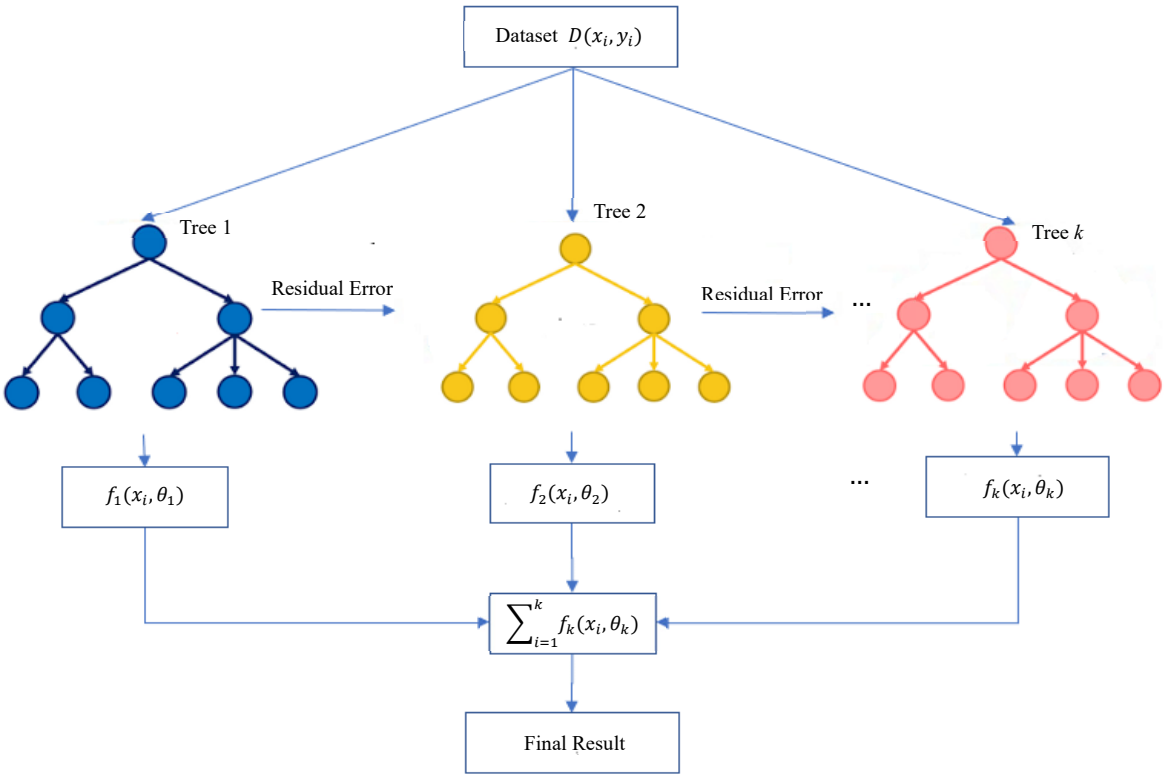

Fig. 1: Flowchart of proposed methodology for chronic kidney disease detection.

Because the XGBoost method incorporates the sparsity awareness concept for each tree, it has a moral aptitude for prediction. Particularly, in case of classification, in which each tree's sparsity awareness method is constructed out of fundamental calculation that target to speed-up computing process [47].

3. Proposed Approach

This section explains the proposed approach. It aims to build an effective and robust classification model to detect the CKD from not-CKD through the input features of patients. The research approach consists of three main stages given in Figure 2 and described in the following subsections. The data input to the approach is the CKD dataset and the output is a classification result that can be CKD or not-CKD. The dataset used in this approach was compiled in 2015 over a two-month period from CKD patients at Apollo Hospital in India. It is publically available for researchers and can be found in the Chronic Kidney Disease dataset in the University of California, Irvine (UCI) data repository [48]. There are 400 observations or records in this dataset, and they have missing and noisy values. From the 400 records, 150 records of patients without CKD and 250 records with CKD are included in the dataset. As a result, 62.5% of each class have CKD, compared to 37.5% who do not have. These observations span a range of ages, from 2 to 90. The CKD dataset contains 24 features, including 11 numerical and 13 categorical features, as shown in Table I. The 25th feature denotes the classification or state of CKD.

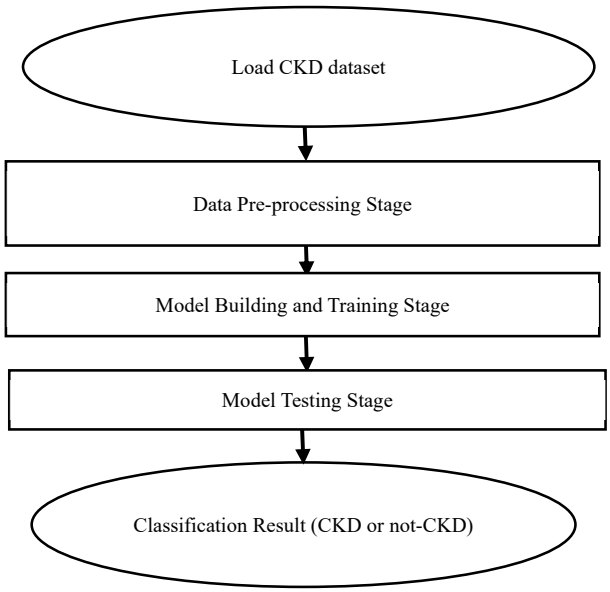

Fig. 2: Flowchart of proposed approach for chronic kidney disease detection.

Table 5: A description of the CKD dataset attributes.

Attributes- (Meaning)	***Value***
age- (The age)	Numerical
bp- (The blood pressure)	Numerical
al- (The albumin)	Numerical (0,1,2,3,4,5)
su- (Sugar)	Numerical (0,1,2,3,4,5)
pcc- (Pus cells clumps)	Categorical (Not-Present, Present)
pc- (Pus cell)	Categorical (Abnormal, Normal)
rbc- (Red blood cells)	Categorical (Abnormal, Normal)
ba- (Bacteria)	Categorical (Not-Present, Present)
bu- (Blood Urea)	Numerical Values in mgs/dl
bgr- (Blood Glucose Random)	Numerical Values in mgs/dl
sc- (Serum creatinine)	Numerical
hemo- (Hemoglobin)	Numerical Values in gms
pot- (Potassium)	Numerical Values in mEq/L
sod- (Sodium)	Numerical Values in mEq/L
pcv- (Packed Cell Volume)	Numerical
rc- (Red blood cell count)	Numerical
wc- (White blood cell count)	Numerical
htn- (Hypertension)	Categorical (No, Yes)
cad- (Coronary Artery Disease)	Categorical (No, Yes)
dm- (Diabetes Mellitus)	Categorical (No, Yes)
appet- (Appetite)	Categorical (Poor, Good)
ane- (Anemia)	Categorical (No, Yes)
pe- (Pedal Edema)	Categorical (No, Yes)
class- (Classification)	Categorical (CKD, not-CKD)

3.1 Data Pre-processing Stage

Data preprocessing is a technique that may be used to transform unclean data into a clean dataset. It is the basic step to train every machine learning classifier algorithm. This method completes tasks like handling missing values, converting it to binary data, and standardizing the dataset. Rescaling is used to scale the dataset when the set of attributes had scales that varied. The mean and most frequent values are used respectively for replacing the missing numerical and categorical values. The categorical values have been transformed into zero and one using the binary transformation. Every attribute's value is regarded as either one for values above the threshold or zero for values below the threshold. Each attribute must have a mean of zero and a standard deviation of one according to the standardized method.

3.2 Model Building and Training Stage

This subsection gives an explanation about model building and training phase. The official XGBoost Python library is used to build the model and initialized its parameters with their default values. After that, the built model is trained using two techniques. The first technique is holdout in which the dataset is divided into 40% for training and 60% for testing. The second technique is a 10-fold cross-validation in which the dataset is divided into 10 parts. Every part is used for testing and the other nine parts are used for training the model. The output of this phase is a trained XGBoost that can be ready for testing and deployment.

3.3 Model Testing Stage

In this stage, the model trained on 40% of the dataset using holdout technique will be tested on the remaining 60% of the dataset. Also, the model trained using a 10-fold cross-validation technique will be tested on a different subset from 10 subsets divided from the dataset for 10 times. The classification results of holdout and 10-fold cross-validation techniques are used to evaluate the proposed approach.

4. Results and Discussion

This section demonstrates the evaluation results of proposed approach to show its performance to detect the CKD from input data features. The results of holdout technique will be given first; then, the results of 10-fold cross-validation technique. The experiments are implemented using a Python programming language. A set of evaluation measures such as recall, precision, accuracy, and F-score are calculated from the classification results. These measures are obtained using the equations listed below:

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

$$F1\text{-}score = 2 * \left(\frac{Recall * Precision}{Recall + Precision}\right) \tag{5}$$

, where TP, FP, FN, and TN are the true positive, false positive, false negative, and true negative cases, respectively.

The classification confusion matrix is used to get the number of true and false positive and negative classified inputs. The results of other evaluation measures are also computed from the confusion matrix and compared with the results of current related work developed for CKD detection on the same dataset. The mean and standard deviation of area under curve (AUC) measure is employed to evaluate the classification results of the 10-fold cross-validation technique.

The number of training and test instances for holdout technique is shown in Figure 3. It can be seen that the number of instances in the training set for CKD and Not-CKD is imbalanced. That means the model should be able to be effective and robust against imbalanced class problem.

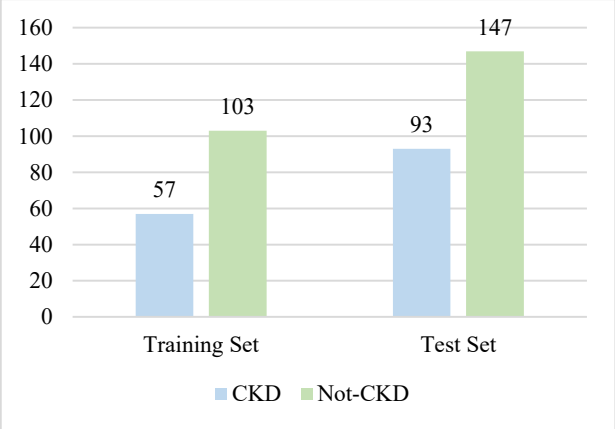

Fig. 3: The number of training and test instances for holdout technique.

After applying the trained XGBoost model on test set, the number of instances that are correctly classified is presented in Figure 4.

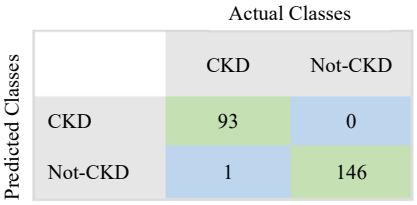

Fig. 4: Confusion matrix for test classification using XGBoost model.

From the confusion matrix output in Figure 4, Table 2 exhibits the evaluation results of the other measures. It illustrates that the model's accuracy attains 99.6% and 0.996 for weighted average recall, precision, and F1-score measures. Also, we can see that the model achieves 1.000 of recall for classifying the CKD test instances and 1.000 for classifying the Not-CKD test instances.

Table 2: Results of evaluation measures for XGBoost model

Class Name	*Recall*	*Precision*	*F-score*	*Accuracy*
CKD	1.000	0.989	0.995	99.6%
Not-CKD	0.993	1.000	0.997	
Weighted Avg.	0.996	0.996	0.996	

To certify the achieved performance of the proposed approach, the average mean and average standard deviation of the 10-fold test examples are computed in Table 3. It can be shown that the model reaches 99.902% of average mean AUC and 0.00710 of average standard deviation AUC. The high value of the average mean AUC confirms that the model has a high performance to differentiate between the negative and positive classes.

Table 3: Results of mean AUC measure for test examples using a 10-fold cross-validation technique

Fold Number	*Test-AUC-mean*	*Test-AUC-std*
1	99.155%	0.02956
2	99.903%	0.01958
3	99.966%	0.01417
4	99.995%	0.00414
5	99.999%	0.00207
6	99.999%	0.00150
7	100.000%	0.00000
8	100.000%	0.00000
9	100.000%	0.00000
10	100.000%	0.00000
Average	99.902%	0.00710

For more analysis, the obtained accuracy result is compared with the current related studies. Figure 5 shows a comparative analysis of the proposed approach against the performance of related work methods.

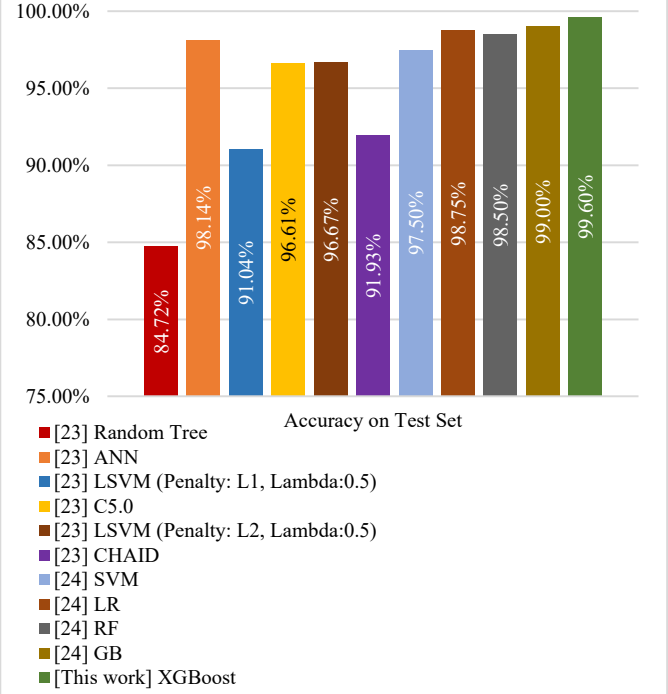

Fig. 5. Accuracy results of XGBoost model compared to the other models in the recent related work.

From the comparison in Figure 5, we can see that the proposed approach using XGBoost model achieves a high accuracy result (99.6%) compared with the classification models in the recent related work. The XGBoost works well due to its ability in-build features to manipulate missing data and its regularization capability to handle unbalanced datasets.

5. Conclusion and Future Work

Detecting chronic kidney disease (CKD) from patients' data using classification algorithms has become a significant task for aiding the doctors to identify the disease early and ensure that patients receive an effective course of treatment. In this paper, a diagnosis approach for detecting CKD using extreme gradient boosting (XGBoost) model is developed to assist the experts for exploring preventive measures early. The approach is trained and evaluated on a public dataset acquired from the UCI repository. The dataset consisted of 24 features for 400 patients. The mean and most frequent values are used respectively for replacing the missing values. The experimental results are conducted using a 10-fold cross-validation and holdout test techniques. Also, a number of evaluation measures are used for obtaining the performance results. The results exposed that the XGBoost model of the proposed approach achieves a competitive high result compared with the recent related work on the same dataset. It attained 99.9% of AUC mean for the 10-fold cross-validation test and 99.6 of accuracy for 60% holdout test from the dataset to diagnosis the CKD effectively. In a future work, a large size dataset will be collected to provide the applicability of deep learning models for CKD detection. Also, they will be compared with this research approach in terms of accuracy and F1-score.

Acknowledgment

"This research was supported by the Researchers Supporting Project number (RSP2022R493), King Saud University, Riyadh, Saudi Arabia."

References

[1] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. Health Soc. Behav.*, vol.24, no. 4, pp. 385-396, 1983.

[2] Z. Rayan, M. Alfonse, and A. B. M. Salem, "Machine Learning Approaches in Smart Health," *Procedia Computer Science*, vol. 154, pp. 361-368, 2018.

[3] P. A. Mccullough, V. Agrawal, E. Danielewicz, and G. S. Abela, "Accelerated atherosclerotic calcification and Mönckeberg's sclerosis: a continuum of advanced vascular pathology in chronic kidney disease," *Clin JAm Soc Nephrol*, vol. 3, pp. 1585-1598, 2008.

[4] "Chronic Kidney Disease," | Harrison's Principles of Internal Medicine, 20e | Access Medicine | McGraw-HillMedical. [Online].Available:https://accessmedicine.mhmedical.com/content.aspx? bookid=2129&sectionid=186950702. [Accessed: 25-Aug-2022].

[5] "Chronic Kidney Disease Clinical Presentation: History, Physical Examination," [Online]. Available:https://emedicine.medscape.com/article/238798-clinical. [Accessed: 25-Aug-2022].

[6] R. Vanholder, S. Van Laecke, G. Glorieux, F. Verbeke, E. Castillo-Rodriguez, and A. Ortiz, "Deleting death and dialysis: Conservative care of cardio-vascular risk and kidney function loss in chronic kidney disease (CKD)," *Toxins*, vol. 10, no. 6, pp. 237-300, 2018.

[7] V. Jha, G. Garcia-Garcia, K. Iseki, Z. Li, S. Naicker, B. Plattner, R. Saran, A. Y. M. Wang, and C. W. Yang, "Chronic kidney disease: Global dimension and perspectives," *The Lancet*, vol. 382, no. 9888, 2013, 260–272.

[8] M. Naghavi, H. Wang, R. Lozano, A. Davis, X. Liang, and M. Zhou, "GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 385, no. 9963, pp. 117-171, 2015.

[9] A. Schieppati, and G. Remuzzi, "Chronic renal diseases as a public health problem: Epidemiology, social, and economic implications," *Kidney International*, vol. 68, no. 98, pp. S7-S10, 2005.

[10] J. W. Stanifer, A. Muiru, T. H.Jafar, and U. D. Patel, "Chronic kidney disease in low- and middle-income countries," *Nephrology Dialysis Transplantation*, vol. 31, no. 6, pp. 868-874, 2016.

[11] K. T. Mills, Y. Xu, W. Zhang, J. D. Bundy, C. S. Chen, T. N. Kelly, J. Chen, and J. He, "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010," *Kidney Int.*, vol. 88, no. 5, pp. 950-957, 2015.

[12] N.R. Hill, S. T. Fatoba, J. L. Oke, J. A. Hirst, C.A. O'Callaghan, D. S. Lasserson, and F. R. Hobbs, "Global prevalence of chronic kidney disease–a systematic review and meta-analysis," *PloS one*, vol. *11*, no. 7, p.e0158765, 2016.

[13] Q. L. Zhang, and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," *BMC public health*, vol. 8, no. 1, pp.1-13, 2008.

[14] M. N. Huda, K. S. Alam, and Harun-Ar-Rashid, "Prevalence of chronic kidney disease and its association with risk factors in disadvantageous population," *Int. J. Nephrol.*, vol. 2012, pp. 1-7, 2012.

[15] C. Tang, J. Ji, Y. Tang, S. Gao, Z. Tang, and Y. Todo, "A novel machine learning technique for computer-aided diagnosis," *Engineering Applications of Artificial Intelligence*, vol. *92*, p.103627, 2020.

[16] M. M. Nishat, F. Faisal, M. A. Mahbub, M. H. Mahbub, S. Islam, and M. A. Hoque, "Performance Assessment of Different Machine Learning Algorithms in predicting Diabetes Mellitus," *Biosc. Biotech. Res. Comm*., vol. 14, no. 1, pp. 74-82, 2021.

[17] F. Faisal, and M. M. Nishat, "An Investigation for Enhancing Registration Performance with Brain Atlas by Novel Image Inpainting Technique using Dice and Jaccard Score on Multiple Sclerosis (MS) Tissue," *Biomedical and Pharmacology Journal*, vol. 12, no. 3, pp. 1249-1262, 2019.

[18] M. R. Farazi, F. Faisal, Z. Zaman, and S. Farhan, "Inpainting multiple sclerosis lesions for improving registration performance with brain atlas," In *2016International Conference on Medical Engineering, Health Informatics and Technology (MediTec,)* pp. 1-6. IEEE, 2016.

[19] S. Ashraf, N. Rehman, H. AlSalman, and A. H. Gumaei, "A decision-making framework using q-rung orthopair probabilistic hesitant fuzzy rough aggregation information for the drug selection to treat COVID-19," *Complexity*, vol. 2022, 2022.

[20] H. Ullah, M. B. Bin Heyat, H. AlSalman, H. M. Khan, F. Akhtar, A. Gumaei, A. Mehdi, A. Y. Muaad, M. S. Islam, A. Ali, and Y. Bu, "An Effective and Lightweight Deep Electrocardiography Arrhythmia Recognition Model Using Novel Special and Native Structural Regularization Techniques on Cardiac Signal," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[21] J. Iqbal, M. Adnan, Y. Khan, H. AlSalman, S. Hussain, S. S. Ullah, and A. Gumaei, "Designing a healthcare-enabled software-defined wireless body area network architecture for secure medical data and efficient diagnosis," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[22] G. Battineni, N. Chintalapudi, and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. *6*, no. 23, pp.e4-e4, 2020.

[23] B. Sahu, S. Mohanty, and S. Rout, "A hybrid approach for breast cancer classification and diagnosis," *EAI Endorsed Transactions on Scalable Information Systems*, vol. *6*, no. 20, pp.e2-e2, 2019.

[24] J. S. Suri, M. Bhagawati, S. Paul, A. Protogeron, P.P. Sfikakis, G. D. Kitas, N. N. Khanna, Z. Ruzsa, A. M. Sharma, S. Saxena, and G. Faa, "Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review," *Computers in biology and medicine*, p.105204, 2022.

[25] C. Chakraborty, and A. N. Abougreen, "Intelligent Internet of Things and Advanced Machine Learning Techniques for COVID-19," *EAI Endorsed Transactions on Pervasive Health and Technology,* vol. 7, no. 26, p.e1, 2021.

[26] A. Gumaei, R. Sammouda, M. Al-Rakhami, H. AlSalman, and A. El-Zaart, "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression," *Health Informatics Journal*, vol. 27, no. 1, p.1460458221989402, 2021.

[27] R. Sammouda, A. Gumaei, and A. El-Zaart, "Intelligent computer-aided prostate cancer diagnosis systems: state-of-the-art and future directions," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[28] S. Sharma, and G. Singh, " Diagnosis of cardiac arrhythmia using Swarm-intelligence based Metaheuristic Techniques: A comparative analysis, " *EAI Endorsed Transactions on Pervasive Health and Technology, vol.* 6, no. 23, 2020.

[29] H. K. Bharadwaj, A. Agarwal, V. Chamola, N. R. Lakkaniga, V. Hassija, M. Guizani, and B. Sikdar, "A review on the role of machine learning in enabling IoT based healthcare applications," *IEEE Access*, vol. 9, no. 2021, pp. 38859-38890, 2021.

[30] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "Performance comparison of some classifiers on chronic kidney disease data," In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, pp. 1-4. IEEE, 2018.

[31] W. H. S. D. Gunarathne, K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," In *2017 IEEE17th International Conference on Bioinformatics and Bioengineering (BIBE),* pp. 291-296. IEEE, 2017.

[32] N. Tazin, S. A. Sabab, and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec),* pp. 1-6, IEEE, 2016.

[33] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *Journal of medical systems,* vol. 41, no. 4, pp. 41-55, 2017.

[34] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC),* pp. 679-684. IEEE, 2019.

[35] M. D. Başar, P. Sarı, N. Kılıç, and A. Akan, "Detection of chronic kidney disease by using Adaboost ensemble learning approach," In *2016 24th Signal Processing and Communication Application Conference (SIU),* pp. 773-776.IEEE, 2016.

[36] A. F. Indriani, and M. A. Muslim, "SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 10, no. 02, pp. 119-127, 2019.

[37] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasiński, Ł. Jasiński, R. Gono, E. Jasińska, and V. Bolshev, "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, no. 2021, pp. 17312-17334, 2021.

[38] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," *IEEE Access,* vol. 8, no. 2020, pp. 25407–25419, 2020.

[39] M. M. Nishat, F. Faisal, R. R. Dip, M. F. Shikder, R. Ahsan, M. A. A. R. Asif, and M. H. Udoy, "Performance Investigation of Different Boosting Algorithms in Predicting Chronic Kidney Disease," In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI),* pp. 1-5. IEEE, 2020.

[40] M. Almasoud, and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *International Journal of Soft Computing and Its Applications*, vol. 10, no. 8, 2019.

[41] A. Andriyashin, "Financial Applications of Classification and Regression Trees," *Master's thesis, Humbolt University*, Berlin, 2005.

[42] E. M. Oliveira and F. L. C. Oliveira, "Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods." *Energy* 144 (2018): 776-788.

[43] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119-139, 1997.

[44] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol.1, no. 1, pp. 14-23, 2011.

[45] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.

[46] Z. Xu, L. Yan, and M. Wang, "Complex Production Process Prediction Model Based On EMD-XGBOOST-RLSE," *The 9th International Conference on Modelling, Identification and Control (ICMIC)*, pp.940 – 947, 2017.

[47] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," I*nternational Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 785-794, 2016.

[48] L. Rubini, "Chronic_Kidney_Disease DataSet," UCI *Machine Learning Repository*, 2015. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. [Accessed: 10-Aug-2022].

Bader Fahad Alkhamees received his Bachelor's degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2003. He received his Master's degree in Software Systems from Heriot Watt University, Scotland, United Kingdom, in 2008. He also received his Ph.D. degree in Biomedical Informatics from Rutgers University, New Jersey, United States. Currently, He is an assistant professor at the Information System Department, College of Computer and Information Sciences, King Saud University. His research interests include Biomedical Informatics, Medical Imaging and Diagnosis, Fuzzy Systems, Computer Networks, the Internet of Things, Machine Learning, Cloud and Edge Computing.