DIAGNOSIS OF CHRONIC KIDNEY DISEASE USING EFFECTIVE CLASSIFICATION ALGORITHM

Mr.A.PANDIAN M.E (Ph.D), M.A.Tamil Selvan, A.Udhayaraj, V.Vasantha Kumar, B.Vignesh

ABSTRACT

Around 10% of the adult population globally is afflicted by chronic kidney disease (CKD), one of the top 20 killers worldwide. CKD is a condition that impairs healthy kidney function. Effective preventative methods for the early detection of CKD are needed due to the rise in CKD cases. The uniqueness of this work is in creating a method of diagnosis to identify chronic renal conditions. This study aids professionals in investigating CKD prevention strategies through early identification utilizing machine learning methods. %is study focuses on examining a dataset gathered from 400 patients having 24 characteristics. The missing nominal and numerical data were replaced using the mean and mode statistical analysis methods. Recursive Feature Elimination is used to determine which features are most crucial (RFE). Blood metabolite creatinine is highly associated with glomerular filtration rate (GFR).Although measuring GFR is challenging, the stage of chronic renal disease is first determined by the creatinine number and subsequently GFR (CKD).CKD could be identified by including a creatinine test in standard medical examinations. Creatinine testing is frequently excluded from standard health examinations since it is more expensive and requires more items for a thorough examination. The likelihood of early detection and treatment of CKD will increase if an algorithm is used to evaluate the risk of the condition without the use of a creatinine test on the findings.

KEYWORD

Chronic, creatinine, glomerular,

INTRODUCTION

Due to its high mortality rate, chronic kidney disease (CKD) has attracted a lot of interest. The World Health Organization (WHO) warns that chronic diseases are now a threat to emerging nations [1]. Early-stage CKD kidney disease is curable, while later-stage CKD results in renal failure. Worldwide, chronic kidney

disease claimed the lives of 753 million people in 2016; 336 million of those deaths were male and 417 million were female [2]. Because kidney disease develops gradually and lasts for a long time, it is referred to be a "chronic" condition. Hindawi Journal of Healthcare Engineering, Volume 2021, Article ID 1004767, 10 pages, https://doi.org/10.1155/2021/1004767 extended period of time, which has an impact on how well the urinary system works. The buildup of waste materials in the blood causes various health issues to surface, which are accompanied by a number of symptoms like high and low blood pressure, diabetes, nerve damage, and bone issues, which result in cardiovascular disease. Diabetes, high blood pressure, and cardiovascular disease (CVD) are risk factors for CKD patients [3]. Patients with CKD experience side effects, particularly in the late stages, which weaken the immunological and nervous systems. Patients may be in advanced stages when they are treated in developing nations, necessitating dialysis or kidney\transplantation. GFR, a measure of renal function, is used by medical professionals to diagnose kidney disease. GFR is calculated using data on the patient's age, gender, blood test results, and other conditions [4]. Medical professionals can categorize CKD into five stages based on the GFR number. Table 1 displays the various kidney disease progression phases together with GFR values. Kidney failure can be avoided if chronic renal disease is detected and treated early. The best method to manage chronic kidney disease is to find it early on, as waiting until it's advanced would result in renal failure and the need for ongoing dialysis or a kidney transplant in order to live a normal life. Two medical tests are used in the medical diagnosis of chronic renal disease. are used to identify CKD, either through a urine albumin test or a blood test to examine the glomerular filtrate. There is a need for computer-assisted diagnostics to assist doctors and radiologists in supporting their diagnostic decisions because of the rising number of patients with chronic kidney disease, the dearth of specialists, and the high costs of diagnosis and treatment, especially in developing countries. Machine learning and deep learning techniques have been used in the processes of disease prediction and early disease detection, and these applications have been made in the health sector and medical image processing. Approaches utilising artificial intelligence (ANN) have been fundamental in the early identification of CKD. Algorithms for machine learning are utilized. for the quick identification of CKD. The two most popular technologies are the ANN and SVM algorithms. These technologies offer significant benefits for diagnosing a variety of conditions, including medical conditions. The ANN algorithm is similar to human neurons in that it can generalize and solve new issues (test data) after being properly taught. The SVM

method, however, uses examples and expertise to assign labels to classes. In essence, the SVM algorithm divides the data along a line that maximizes the distance between the class data [6]. Several factors affect renal performance, which produce CKD,\slike diabetes, blood pressure, heart disease, some kind of\sfood, and family history. A few contributing variables to chronic kidney disease are shown in Figure 1. A technique for identifying the stages of CKD using ultrasonography (USG) pictures was presented by Pujari et al. [7]. The system finds instances of fibrosis at various times. In order to identify whether the urinary system is healthy or unhealthy, Ahmed et al. [8] created a fuzzy expert system. The properties of CKD were extracted using a stacked autoencoder model by Khamparia et al. [9] and the resultant class was then classified using Softmax. A genetic algorithm (GA) based on neural networks was proposed by Kim et al. [10] in which the weight vectors were improved using GA to train NN. For CKD diagnosis, the system outperforms conventional neural networks. A model based on neural networks was proposed by Vasquez-Morales et al. [11] to determine whether a person is at risk of getting CKD. A CKD dataset was diagnosed by Almansour et al. [12] using ANN and SVM methods. Accuracy levels for ANN and SVM were 99.75% and 90.75%, respectively. Radial basis function (RBF), multilayer perceptron (MLP), SVM, and probabilistic neural networks (PNN), among other methods, were used by Rady and Anwar [13] to detect CKD.

Table 1: CKD Development Stages

Stage	Description	GFR(ml/min/1.73 m^2)	Treatment stage
1	Normal kidney function	>= 90	Controlling blood pressure
2	Mild kidney damage	60-90	Controlling blood pressure and its risk factor
3	Moderate kidney damage	30-59	Controlling blood pressure and its risk factor

4	Severe kidney damage	15-29	End stage renal failure
5	Established kidney failure	<=15	Treatment chices

MATERIALS AND METHODS

To assess the CKD dataset, a number of experiments were carried out utilising the machine learning algorithms SVM, KNN, decision tree, and random forest. The general organisation of CKD diagnosis in this work is depicted in Figure 2. The missing nominal values were computed using the mode technique during preprocessing, and the missing numerical values were computed using the mean method. %e features of relevance related with the characteristics of importance for CKD diagnosis were identified using the RFE method. For the purpose of diagnosing diseases, classifiers were fed these chosen features. In this work, SVM, KNN, decision trees, and random forests were used as classifiers to diagnose CKD. When categorizing a dataset into having CKD or having a normal kidney, all classifiers produced encouraging results.

DATASET

400 patients from the University of California, Irvine Machine Learning Repository provided data for the CKD dataset [24]. The 24 features in the dataset are broken down into 11 numerical features, 13 categorical characteristics, and class features like "ckd" and "notckd" for classification. Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell clumps, bacteria, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edoema, and anaemia are among the

characteristics. There are two values in the diagnostic class of %e: ckd and notckd. Except for the diagnostic feature, every feature had missing values. Due to the 250 instances of "ckd" in the dataset, it is unbalanced 150 "notckd" cases were reduced by 37.5%, and class by 62.5%.

PREPROCESSING

The dataset needed to be cleaned up in a preprocessing stage because it had outliers and noise. Preprocessing steps included checking for uneven data, normalizing the data, estimating missing values, and removing noise like outliers. When patients are undergoing tests, it is possible for some measures to be missed, leading to missing numbers. 158 occurrences in the dataset were complete, whereas the rest instances had missing values. Ignoring the record is the easiest way to deal with missing values, however it is not appropriate for tiny datasets. Instead of deleting records, we can apply algorithms to compute missing data. One of the statistical measures, such as mean, median, and standard deviation, can be used to calculate the percentage of missing values for numerical features. However, the mode approach, which substitutes the missing value for the most typical value of the nominal feature, can be used to compute the missing values of nominal features the attributes. In this study, the mean approach was used to replace the missing numerical features, and the mode method was used to replace the missing nominal features. The statistical analysis of the dataset is shown in Table 2, including the mean, standard deviation, and the introduction of max and min for the numerical features in the dataset. The statistical analysis of a numerical feature is shown in Table 3. While numerical features can be either separate or continuous, they are values that can be measured.

FEATURES SELECTION

It is necessary to find the significant aspects that have a strong and positive connection with features of importance for disease diagnosis after computing the missing values. A robust diagnostic model cannot be built since the vector characteristics must be extracted to remove unnecessary and useless features for prediction [25]. In this work, the most crucial elements of a prediction were extracted using the RFE approach. Because to its simplicity of usage and setups, as well as its efficiency in picking features in training datasets important to predicting target variables and removing weak features, the Recursive Feature Elimination (RFE) algorithm is particularly well-liked. By identifying high correlation between particular characteristics and the target, the %e RFE approach is used to choose the most important features (labels). Figure 4 displays the most important characteristics, as determined by RFE; it is highlighted that the albumin characteristic has the highest correction (17.99%), featured by 14.34%, followed by packed cell volume (12.91%), and serum creatinine (12.09%). Figure 3 displays RFECV, which depicts the dataset's feature count along with a cross-validated score and illustrates the features that were chosen.

CLASSIFICATION

To create classification templates, novel and comprehensible patterns have been defined using data mining approaches [26]. In order to do classification and regression using supervised and unsupervised learning approaches, models must first be built based on prior analysis [27]. SVM, KNN, decision trees, and random forests are four widely used machine learning algorithms that produce the greatest diagnostic outcomes. Machine learning techniques work to build predictive/classification models through two\stages: the training phase, in which a model is constructed\from a set of training data with the expected outputs, and\the validation stage, which estimates the quality of the\strained models from the validation dataset without the\expected

output. Any algorithm used to address classification problems is supervised.

SUPPORT VECTOR MACHINE CLASSIFIER

The SVM algorithm essentially draws a line dividing the dataset into classes, allowing it to determine which class the test data belongs to classes it falls under. A hyperplane is a decision boundary or %e line. Both linear and nonlinear types can be used with %e method. When the dataset has two classes and is separable, linear SVM is utilized. A nonlinear SVM is used when the dataset cannot be separated, and the algorithm changes the original coordinate region into a separable space. Several hyperplanes may exist, and the best hyperplane is selected using the largest possible margin between data points. A support vector is the dataset that is most closely related to the hyperplane.

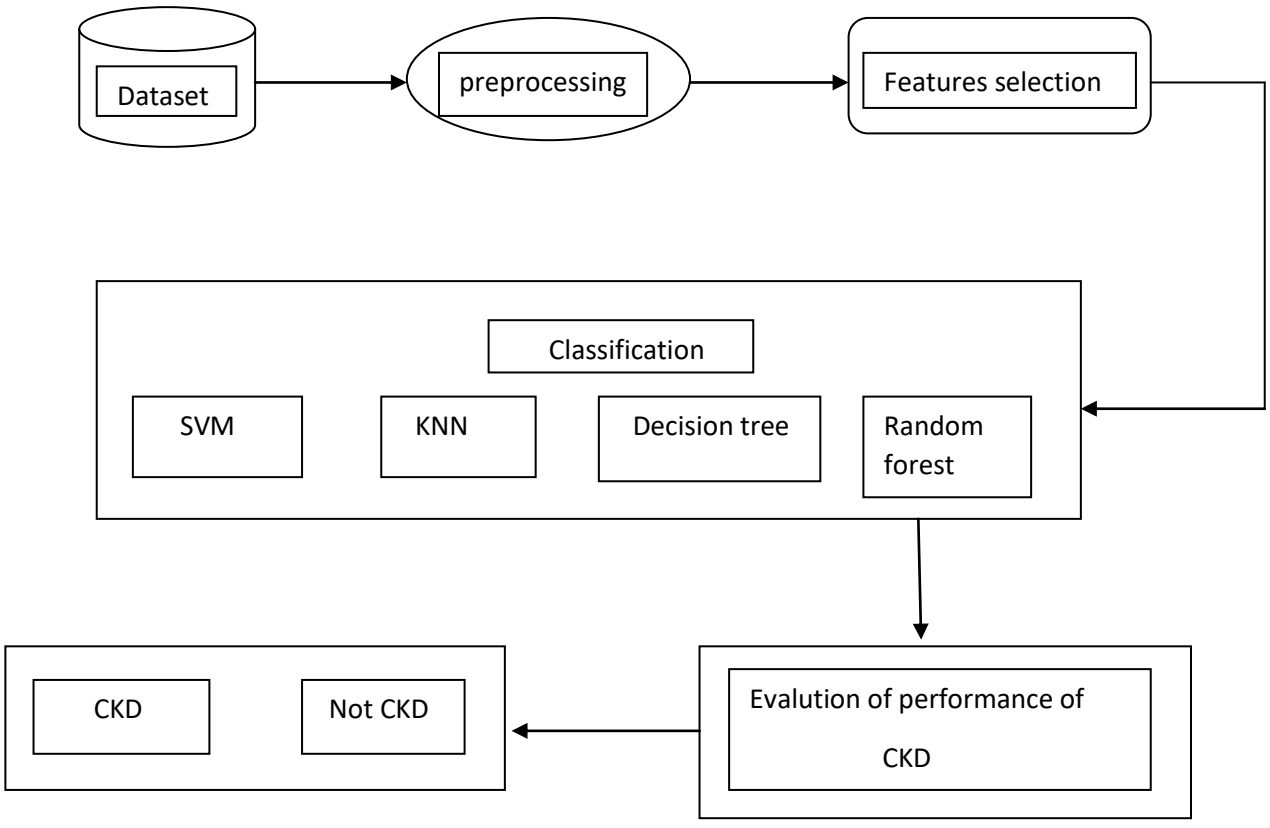

Figure 1: The proposed system for diagnosis of CKD.

$$K(X,X') = \exp(\| X - X' \|^2 / 2 \&^2)$$

where "X X′" 2 denotes the distance between the input features and "X, X′" are input data. A free parameter is.

For the classification of the data, the Radial Basis Function (RBF) was used.

K-NEAREST NEIGHBOR CLASSIFIER

The KNN algorithm divides the set of classes into those that are most similar to the new test point and those that are most similar to the training points. The lazy learning algorithm, also known as the %e KNN algorithm, is a nonparametric algorithm that maintains training data rather than learning from it. The Euclidean distance is used to calculate the distance between the newly added point and the previously stored training points while classifying the new dataset (test data). The class with the most neighbours is assigned to the new point. The nearest neighbour in the features vector was determined using the %e Euclidean distance function (Di).

DECISION TREE CLASSIFIER

The foundation of a decision tree algorithm is a tree structure. The leaf node represents the result, the branches represent the decision rules, the internal nodes represent the features, and the root node represents the whole dataset. A decision node, which has more branches, and a leaf node, which has no additional branches, are both types of nodes found in a decision tree. Choices are made based on the features provided. The algorithm makes a conclusion and progresses to the next node based on

the comparison between the feature in the root node and the feature's record (actual dataset) in a decision tree. Up until it reaches the leaf node, the algorithm compares the features in the second node with the features in the subnodes.

RANDOM FOREST CLASSIFIER

The Random Forest approach utilizes a number of classifiers in an ensemble to enhance model performance and address challenging problems. According to the algorithm's name, it is a classifier that uses decision trees on subsets of the dataset and takes an average to enhance prediction. The random forest method collects predictions from each decision tree and uses the majority vote to decide whether to anticipate the final outcome rather than depending solely on one decision tree for the prediction process. The accuracy increases with the number of trees used, which avoids the overfitting issue. Some trees may predict the correct output while others may not because the technique uses some decision trees to forecast the class of a dataset. Hence, there are two presumptions for a prediction's high accuracy. In order for the algorithm to anticipate accurate results rather than speculative ones, the feature variable must first contain valid values. Second, there should be extremely little correlation between each tree's predictions. Hence, there are two prerequisites for a high degree of prediction accuracy. In order for the algorithm to anticipate accurate results rather than assuming, the feature variable must first contain actual values. Second, there should be a strong correlation between each tree's predictions.

EXPERIMENT ENVIRONMENTAL SETUP

The results of the developing system was presented in this session.

ENVIRONMENT SETUP

Several environments have been used in the development of the system. The environment configuration for the evolving system is displayed.

Table 2: Proposed system's environment setup

Resource	**Specification**
CPU	Core i6 gen6
RAM	12 GB
GPU	4GB
software	Python

Table 3: Dataset Splitting

Dataset	**Numbers**
Training	400 patients
Testing and validation	100 patients

Table 4: Outcomes of diagnosing CKD using four machine learning methods.

classifiers	**KNN**	**SVM**	**Random forest**	**Decision tree**
Accuracy %	99	98	100	100
Precision %	100	95	100	100
Recall %	99	98	100	99.34
F1-score %	99.5	99	100	100

EVALUTION MATRICS

Performance indicators were employed to assess each of the four classifiers' capabilities. Among these measurements is the confusion matrix, which yields the

Calculating the correctly categorised samples (TP and TN) and the wrongly classified samples (FP and FN) yields the accuracy, precision, recall, and F1-score, as illustrated in the following equations [28]:

Recall= (TP/TP+FN)*100%

Precision= (TP/TP+FN)*100%

Accuracy= (TN+TP/TN+TP+FN+FP)*100%

SPLITING DATASET

The dataset was split into 25% for testing and validation and 75% for training. The split data are displayed in Table 3.

RESULTS

All positive samples were accurately identified for 250 samples (TP) by the random forest algorithm, while all negative samples (TN) were correctly classified for 150 samples (TN). While the positive (TP) samples were assessed by the SVM, KNN, and Decision Tree algorithms by 94.74%, 97.37%, and 98.68%, respectively, with an error (TN) of 5.26%, 2.63%, and 1.32%, respectively. The findings produced by the four classifiers are displayed in Table 6. The performance of our suggested system was compared to that of earlier studies in Table 8, where the random forest approach outperformed the other classifiers. Performance evaluation of the system's diagnostic precision across the two datasets. 8 Journal of Healthcare Engineering recall, with a total F1-score of 1. It was followed by the decision tree algorithm, which achieved scores of 99.17%, 100%, 98.68%, and 99.34% for accuracy, precision, recall, and F1-score, respectively. The accuracy, precision, recall, and F1-score of the KNN algorithm were then 98.33%, 100%,

97.37%, and 98.67%, respectively. The final results were 96.67%, 92%, 94.74%, and 97.30% for the SVM accuracy, precision, recall, and F1-score algorithm, respectively. As stated in Table 4, the effectiveness of the suggested methods was assessed through a number of prior similar research. It should be highlighted that the available research have obtained the lowest levels of accuracy, with a range of 96.8%. and While the accuracy of the suggested system with the random forest tree method is 100%, it is only 66.3%. When compared to current systems, it is seen that the proposed has the best outcomes. Among 400 CKD patients, 24 numerical and nominal features were introduced. Several computing approaches were used to solve this problem because some tests for some patients were neglected. The mean approach was used to solve the missing numerical values, and the mode method was utilised to solve the missing nominal values. The association between several variables in Figure 4 is depicted as both positive and negative correlation. For instance, there is a positive association between specific gravity and haemoglobin, packed cell volume, and red blood cell count; between sugar and blood glucose random; and between blood urea and Hemoglobin, red blood cell count, and packed cell volume are all correlated with serum creatinine. Also, there is a negative link, for instance, between albumin and blood urea and haemoglobin, packed cell volume, red blood cell count, and serum creatinine and sodium.

RESULTS AND DISCUSSION

The dataset is randomly split into 25% for testing and validation and 75% for training. The irrelevant subset features were chosen using the %e Recursive Feature Elimination approach, which was provided. Afterwards, classifiers were used to process the selected features in order to diagnose CKD. Table 8 provides an evaluation of the proposed system in comparison to current methods. It should be highlighted that the suggested system has produced encouraging outcomes. In order to prioritise the features and assign a percentage to each feature based on

the correlation with the target feature, we utilised the RFE algorithm to determine the best associations between each feature and the target features. Figure 5 compares the performance of the proposed system to that of existing systems, whose accuracy peaked at a ratio of between 95.84% and 66.3%, and our systems' accuracy ranged from 100% by random forest to 97.3% by SVM.

CONCLUSION

This study shed light on how CKD patients should be diagnosed in order to address their condition and obtain treatment in the earliest stages of the disease. 400 patients made up the dataset, which contained 24 characteristics. The dataset was split among 25% testing and validation and 75% training. The mean and mode statistical measures were used, respectively, to restore missing numerical and nominal values and remove outliers from the dataset. The RFE technique was used to pick the CKD traits that were most highly representative. SVM, KNN, decision trees, and random forests were among the classification algorithms that received input from chosen features. All classifiers' parameters were adjusted to provide the best classification, and as a consequence, all algorithms produced encouraging results.

All other algorithms were outperformed by the random forest approach, which achieved an accuracy, precision, and Recall, and a 100% F1-score for all measures. The empirical results of SVM, KNN, and decision tree algorithms discovered significant values of 96.67%, 98.33%, and 99.17% with respect to accuracy metric after the system was investigated and evaluated by multiclass statistical analysis.

REFERENCES

[1] World Health Organization, Preventing Chronic Disease: A Vital Investment, WHO, Geneva, Switzerland, 2005.

[2] B. Bikbov, N. Perico, and G. Remuzzi, “Disparities in chronic kidney disease prevalence amongmales and females in 195 countries: analysis of the global burden of disease 2016 study,” Nephron, vol. 139, no. 4, pp. 313–318, 2018.

[3] Z. Chen, X. Zhang, and Z. Zhang, “Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models,” International Urology and Nephrology, vol. 48, no. 12, pp. 2069–2075, 2016.

[4] Glomerular Filtration Rate (GFR), National Kidney Foundation, New York, NY, USA, 2020, https://www.kidney.org/ atoz/content/gfr.

[5] T. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, “Soft computing model to predict chronic diseases,” Information Science and Engineering, vol. 36, no. 2, pp. 365–376, 2020.

[6] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” Bioinformatics, vol. 16, no. 10, pp. 906–914, 2000.

[7] R. M. Pujari and V. D. Hajare, “Analysis of ultrasound images for identification of Chronic Kidney Disease stages,” in Proceedings of the 2014 First International Conference on Networks & Soft Computing (ICNSC2014), pp. 380–383, IEEE, Guntur, India, August 2014.

[8] S. Ahmed, M. T. Kabir, N. T. Mahmood, and R. M. Rahman, “Diagnosis of kidney disease using fuzzy expert system,” in Proceedings of the 8th International Conference on Software, Knowledge, Information Management and Applications Journal of Healthcare Engineering 9 (SKIMA 2014), pp. 1–8, IEEE, Dhaka, Bangladesh, December 2014.

[9] A. Khamparia, G. Saini, B. Pandey, S. Tiwari, D. Gupta, and A. Khanna, “KDSAE: chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network,” Multimedia Tools and Applications, vol. 79, no. 47-48, pp. 35425–35440, 2019.

[10] D. -H. Kim and S. -Y. Ye, “Classification of chronic kidney disease in sonography using the GLCM and artificial neural network,” Diagnostics, vol. 11, no. 5, p. 864, 2021.

[11] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, “Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning,” IEEE Access, vol. 7, pp. 152900–152910, 2019.

[12] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, and J. Alhiyafi, “Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study,” Computers in Biology and Medicine, vol. 109, pp. 101–111, 2019.

[13] E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” Informatics in Medicine Unlocked, vol. 15, Article ID 100178, 2019.

[14] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic kidney disease analysis using data mining classification techniques,” in Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), pp. 300–305, IEEE, Noida, India, January 2016.

[15] M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, “Boosted classifier and features selection for enhancing chronic kidney disease diagnose,” in Proceedings of the 2017 5th international conference on cyber and IT service management (CITSM), pp. 1–6, IEEE, Denpasar, Indonesia, August 2017.

[16] E. Avci, S. Karakus, O. Ozmen, and D. Avci, “Performance comparison of some classifiers on chronic kidney disease data,” in Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–4, IEEE, Antalya, Turkey, March 2018.

[17] R. K. Chiu, R. Y. Chen, S. A. Wang, Y. C. Chang, and L. C. Chen, “Intelligent systems developed for the early detection of chronic kidney disease,” Advances in Artificial Neural Systems, vol. 2013, 2013