International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Machine Learning Approach to Chronic Kidney Disease

S. Sravan Kumar Reddy, Dr. Srinivasan Jagannathan, Mr. Suresh
Department of Computer Application, Madanapalle Institute of Technology and Science, Madanapalle, India

Article Info

Publication Issue :
Volume 8, Issue 5
September-October-2022

Page Number : 01-09

Article History
Accepted: 20 Aug 2022
Published: 04 Sep 2022

ABSTRACT

Chronic kidney disease (CKD) is a worldwide health problem that causes significant morbidity and mortality, as well as the onset of other illnesses. People frequently miss CKD because there are no obvious symptoms in the early stages. Early detection of CKD allows patients to receive timely treatment to slow the disease's progression. Machine learning models can successfully assist doctors in achieving this goal due to their rapid and precise identification capabilities. In this paper, we present a machine learning framework for CKD diagnosis. The CKD data set was retrieved from the University of California, Irvine's machine learning repository (UCI). Due to this, it will determine whether a patient has CKD and, if so, whether or not additional medications need to be taken. Models were developed using six machine learning techniques: gradient boosting, logistic regression, adaBoost, random forest, and decision trees. The most accurate machine learning model was random forest. We proposed an integrated model that combines logistic regression and random forest using perceptron, best accuracy, by examining the errors produced by the existing models. We therefore hypothesised that this methodology might be applicable to clinical data for disease diagnosis that is more complex.

Keywords : Logistic Regression, AdaBoost, Random Forest, Decision Tree, Gradient Boosting.

I. INTRODUCTION

Nearly all of the world's population is impacted by CHRONIC kidney disease (CKD), a problem of global public health. Chinese CKD prevalence rates as a percentage and American prevalence rates as a range. The general adult population of Mexico has reached this percentage, according to another study. A slow decline in renal function, which ultimately results in a total loss of renal function, is a hallmark of this illness. In its early stages, CKD doesn't have any overt symptoms. As a result, until the kidney starts to lose function, the disease might not be discovered. Additionally, the global effects of CKD on the human body include high rates of morbidity and mortality. Cardiovascular disease may become more likely as a result. A pathologic syndrome that progresses and cannot be cured is CKD. Therefore, it is very important to predict and diagnose CKD early on because doing so may allow patients to receive treatment in a timely manner, slowing the disease's progression. Machine learning is the process of using

a computer programme to calculate and infer relevant data, determine the traits of a given pattern, and perform other related tasks. It may be a promising method to diagnose CKD because this technology can produce accurate and affordable diagnoses of diseases. With the advancement of information technology, it has taken on a new type of medical tool, and the rapid development of the electronic health record has opened up a wide range of potential applications. Machine learning has already been applied in the medical industry to diagnose different diseases, examine relevant disease-related factors, and determine the health of the human body. For instance, models created by machine learning algorithms have been applied to the diagnosis of cancer, acute kidney injury, diabetes, retinopathy, heart disease, and other diseases. The regression, tree, probability, decision surface, and neural network algorithms in these models were frequently successful. To identify renal morphologic changes in the context of CKD diagnosis, Hodnel and colleagues used image registration. By utilising extensive CKD data and a neural network classifier, Vasquez-Morales et al. were able to demonstrate the model's accuracy using test data. Additionally, the majority of earlier studies made use of the CKD data set, which was obtained from the UCI machine learning repository. Support vector machine (SVM), gradient boosting classifier, and soft independent modelling of class analogy were used by Chen et al. to diagnose CKD. Additionally, they used fuzzy rule-building expert systems, fuzzy optimal associative memories, and partial least squares discriminant analysis to diagnose CKD, and the range of accuracy in those models allowed for successful results in the diagnosis of CKD in their studies. The diagnostic categories of the samples determine how the mean imputation method is used to fill in the missing values in the models mentioned above. Since the diagnostic outcomes of the samples are unknown, their method is therefore not applicable. Before diagnosing, patients actually have a chance to miss some measurements for a variety of reasons. In addition, data derived using mean imputation for categorical variables with missing values may deviate greatly from the true values. For instance, we might set the categories for variables with only two categories to 0 and 1, but the mean of the variables might range from 0 to 1. The proposed models' computational costs were reduced through feature selection, and the range of accuracy in those models was increased, according to Polat et alSVM .'s technology. The missing values were filled in by J. Aljaaf et al. using novel multiple imputation before accuracy was attained by a neural network. In order to diagnose CKD, Subas et al. employed ANN, SVM, Gradient Boosting, decision tree, and random forest (RF), with the RF achieving an accuracy in the models developed by Boukenze et al., while MLP attained the highest accuracy. Studies primarily aim to create models and produce the best possible outcomes. On the other hand, no feature selection technology is used to choose predictors, nor is a thorough procedure for filling in the missing values described in detail. In order to diagnose CKD and assess the accuracy of the models, Almansour et al. used SVM and neural networks, respectively. Zero was used to fill in the missing values in the models created by Gunarathne et al., and decision forest produced the best results in terms of accuracy.

II. RELATED WORKS

Prediction of Chronic Kidney Disease - A Machine Learning Perspective: Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural

network, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority oversampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%. Along with machine learning models one deep neural network has been applied on the same dataset and it has been noted that deep neural network achieved the highest accuracy of 99.6%.

Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm: Chronic Kidney Disease is one among the noncontagious illnesses that affect most of the individual in the world. The main factors of risk for the Chronic Kidney Disease are Diabetes, Heart Ailment, and Hypertension. The Chronic Kidney Disease shows no symptoms in the early stages and most of the cases are diagnosed in the advanced stage. This leads to delayed treatment to the patient which may be fatal. Machine learning technique provides an efficient way in the prediction of Chronic Kidney Disease at the earliest stage. In this paper, four ensemble algorithms are used to diagnose the patient with Chronic Kidney Disease at the earlier stages. The machine learning models are evaluated based on seven performance metrics including Accuracy, Sensitivity, Specificity, F1-Score, and Mathew Correlation Coefficient. Based on the evaluation the AdaBoost and Random Forest performed the best in terms of accuracy, precision, Sensitivity compared to Gradient Boosting and Bagging. The AdaBoost and Random Forest also showed the Mathew Correlation Coefficient and Area Under the curve scores of 100%. The machine learning model proposed in this paper will provide an efficient way to prevent Chronic Kidney diseases by enabling the medical practitioners to diagnose the disease at an early stage.

Early Detection and Prevention of Chronic Kidney Disease: In the age group ranging from 65 – 74 worldwide, it is estimated that one in five men, and one in four women, have Chronic Kidney Disorder (CKD). 10% of the population worldwide is affected by (CKD), and millions die each year due to lack of access to affordable treatment. A protein present in urine, persistent proteinuria is a key indicator for the presence of CKD. Early detection can help prevent progression of kidney disease to kidney failure. This detection and subsequent prevention can be achieved by applying Data Mining techniques on patient information to predict the occurrence of Chronic Kidney Disease. In this research paper, a Data Mining algorithm, Boruta analysis is performed to extrapolate the factors which can fortify the chances of a patient having CKD. This analysis covers statistic data along with historic and medical details. The dataset has been obtained from UCI source which contains data of 400 samples from the southern part of India with their ages ranging between 2-90 years. Making a decision concerning the seriousness of given factors,

an estimate can be drawn with respect to the same. In Australia, treatment for all current and new cases of kidney failure through 2020 will cost an estimated $12 billion. Such an algorithm can help many individuals overall who may experience the ill effects of such affliction in their lifetime. Boruta Analysis, being freely available helps in medical diagnosis which can be otherwise expensive. It makes the diagnosis economical as well as faster for the patients.

Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms: One major cause of death and disability is CKD. In 1990, it was the 27th main reason, and in 2010, it moved up to the 18th main reason. 2013 saw the death of nearly 1 million people. Nevertheless, CKD still has an impact on residents of developing nations. We investigated CKD patient data and put forth a system that can be used to gauge the likelihood of developing CKD. 455 patients' data were used. Both the real-time dataset from Khulna City Medical College and the online data set from UCI Machine Learning Repository are used in this study. As a high-level interpreted programming language, Python was used by us to create our system. We used Random forest and ANN after training the data with a 10-fold CV. ANN and Random Forest both achieve accuracy of 94.5% and 97.12%, respectively. With the aid of this system, chronic kidney diseases will be more likely to surface early.

Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm: Chronic Kidney disease (CKD), a slow and latediagnosed disease, is one of the most important problems of mortality rate in the medical sector nowadays. Based on this critical issue, a significant number of men and women are now suffering due to the lack of early screening systems and appropriate care each year. However, patients' lives can be saved with the fast detection of disease in the earliest stage. In addition, the evaluation process of machine learning algorithm can detect the stage of this deadly disease much quicker with a reliable dataset. In this paper, the overall study has been implemented based on four reliable approaches, such as Support Vector Machine (henceforth SVM), AdaBoost (henceforth AB), Linear Discriminant Analysis (henceforth LDA), and Gradient Boosting (henceforth GB) to get highly accurate results of prediction. These algorithms are implemented on an online dataset of UCI machine learning repository. The highest predictable accuracy is obtained from Gradient Boosting (GB) Classifiers which is about to 99.80% accuracy. Later, different performance evaluation metrics have also been displayed to show appropriate outcomes. To end with, the most efficient and optimized algorithms for the proposed job can be selected depending on these benchmarks.

Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform: Chronic Kidney disease is a severe lifelong condition caused either by renal disease or by impaired functions of the kidneys. In the present area of research, Kidney cancer is one of the deadliest and crucial importance for the survival of the patients ' diagnosis and classification. Early diagnosis and proper therapy can stop or delay the development of this chronic disease into the final stage where dialysis or renal transplantation is the only way of saving the life of the patient. The development of automated tools to accurately identify subtypes of kidney cancer is, therefore, an urgent challenge in the recent past. In this paper, to examine the ability of various deep learning methods an Adaptive hybridized Deep Convolutional Neural Network (AHDCNN) has been proposed for the early detection of Kidney disease efficiently and effectively. Classification technology efficiency depends on the role of the data set. To enhance the accuracy of the classification system by reducing the feature dimension an algorithm model has been developed using CNN. These high-level properties help to build a supervised tissue classifier that discriminates between the two types of tissue. The experimental process on the Internet of medical

things platform (IoMT) concludes, with the aid of predictive analytics, that advances in machine learning which provides a promising framework for the recognition of intelligent solutions to prove their predictive capability beyond the field of kidney disease.

III. PROPOSED SYSTEM

In this section, a number of machine learning models are proposed. To create the classifiers, various machine learning algorithms first used data samples to diagnose the algorithms. These models were evaluated, and the ones that performed the best were chosen as potential candidates. The component models were found by investigating their errors in judgement. Then, to achieve better performance, an integrated model was established.

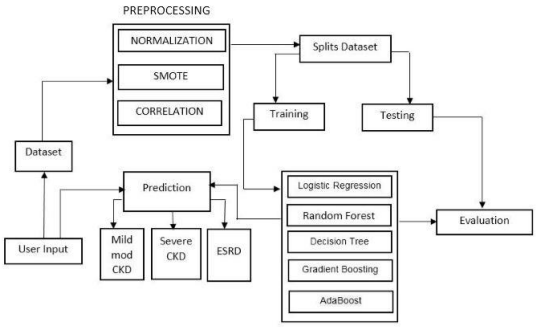


Figure 1: Block diagram

IV. IMPLEMENTATION

The algorithms listed below were used to complete the project.

1. Logistic Regression:

One of the most well-known machine learning algorithms, under the category of supervised learning, is logistic regression. A set of independent variables are used to predict the categorical dependent variable using this method. An output of a dependent categorical variable is predicted by logistic regression. A discrete or categorical value must therefore be the result. It can be either Yes or No, 0 or 1, true or false, etc., but rather than providing an exact value between 0 and 1, it provides probabilistic values that are in the range of 0 and 1. Since they are used differently, linear regression and logistic regression are very similar. In order to solve regression problems, one must use linear regression; however, classification problems require the use of logistic regression. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function. There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So the company wanted to check how many users from the dataset, wants to purchase the car. For this problem, we will build a Machine Learning model using the Logistic regression algorithm. The dataset is shown in the below image. In this problem, we will predict the purchased variable by using age and salary.

2. Random Forest:

A machine learning method called a random forest is employed to address classification and regression issues. It makes use of ensemble learning, a technique that combines a number of classifiers to offer answers to challenging issues.

Deterministic trees make up a random forest algorithm. Through bagging or bootstrap aggregation, the random forest algorithm trains the "forest" it

creates. Machine learning algorithms' accuracy is increased by bagging, an ensemble meta-algorithm.
Based on the decision trees' predictions, the (random forest) algorithm determines the result. The output from various trees is averaged or averaged to make predictions. The precision of the result improves with more trees.
The decision tree algorithm's drawbacks are removed by a random forest. It improves precision and decreases overfitting of datasets. Without many configurations in packages, it generates predictions (like Scikit-learn).
Features:

- Random forest algorithms works more accurate than decision tree algorithm.
- Random forest offers a practical means of dealing with missing data.
- Random forest is capable of producing an accurate prediction without hyper-parameter tuning.
- It fixes the overfitting problem with decision trees.
- A subset of features is randomly chosen at each node's splitting point in every random forest tree.

An algorithm using a random forest has decision trees as its foundation. Using a structure resembling a tree, decision trees are a type of decision support method. We will learn how random forest algorithms operate by first reviewing decision trees.
The three parts of a decision tree are the decision node, the leaf node, and the root node. A training dataset is separated into branches by a decision tree algorithm, which then divides those branches into additional branches. Up until a leaf node is reached, this sequence keeps going. You can't separate the leaf node any more.

3. AdaBoost:

The Boosting technique known as AdaBoost algorithm, also known as Adaptive Boosting, is used as an Ensemble Method in machine learning. The weights are redistributed to each instance, with higher weights being given to instances that were incorrectly classified, hence the name "adaptive boosting." For supervised learning, boosting is used to lower bias and variance. It operates under the premise that students advance in stages. Each learner after the first is developed from a previous learner, with the exception of the first. Simply put, weak students are transformed into strong ones. Similar in concept to boosting, the AdaBoost algorithm differs slightly from it. Let's go over this distinction in more detail.

Let's first talk about how boosting functions. During the data training phase, 'n' decision trees are created. The incorrectly classified record in the first model is given priority as the first decision tree or model is constructed. For the second model, only these records are sent as input. The procedure continues until we decide how many base learners to create. Remember that all boosting techniques permit record repetition.

The first model's creation process is depicted in this figure, along with the algorithm's identification of first model errors. The wrongly classified record serves as the input for the following model. Until the required condition is satisfied, this process is repeated. The figure shows that 'n' different models were created using the mistakes from the previous model. This is the process of boosting. It is possible to refer to the models 1, 2, 3, and N as decision trees because they are individual models. The same principles underlie all different boosting models. The AdaBoost algorithm will be simple to understand now that we are aware of the boosting principle. Enter the workings of AdaBoost. The algorithm produces a 'n' number of trees when the random forest is used. By combining a start node with several leaf nodes, it creates proper trees. There is no set depth in a random forest, although some trees may be larger than others. However, the AdaBoost algorithm only creates the Stump node, which has two leaves.

4. Decision Tree

The real world is full of analogies for trees, and it turns out that these analogies have influenced a large portion of machine learning, including both classification and regression. A decision tree is a visual and explicit representation of decisions and decision-making in decision analysis. It employs a decision-tree structure, as suggested by the name. Despite being a frequently employed tool in data mining for determining a plan of action to accomplish a specific objective.

Drawn inverted, with the root at the top, is a decision tree. In the left image, a condition or internal node, based on which the tree divides into branches or edges, is represented by the bold text in black. The decision/leaf, in this case whether the passenger survived or not, is at the end of the branch that doesn't split any longer and is shown as red and green text, respectively.

You cannot ignore the simplicity of this algorithm, even though a real dataset will have many more features and this will only be one branch in a much larger tree. Relations can be easily seen, and the feature importance is obvious. The above tree is referred to as a "Classification tree" because its goal is to categorise passengers as having survived or having passed away. This methodology is more commonly referred to as "learning decision tree from data." Regression trees are modelled similarly, but instead of predicting discrete values like house prices, they forecast continuous values. CART, which stands for Classification and Regression Trees, is the general name for Decision Tree algorithms.

So, what exactly is happening in the background? Selecting the right features, establishing the right conditions for splitting, and determining when to stop are all important aspects of growing a tree. A tree will need to be pruned in order to look beautiful because it typically grows at random. Start by discussing a splitting method that is frequently used.

5. Gradient Boosting:

One of the strongest algorithms in the machine learning space is the gradient boosting algorithm. Bias error and variance error are the two broad categories into which errors in machine learning algorithms can be grouped. It is used to reduce the model's bias error because gradient boosting is one of the boosting algorithms.

Contrary to the Adaboosting algorithm, we are not able to identify the base estimator in the gradient boosting algorithm. The Gradient Boost algorithm uses Decision Stump as its default base estimator, which is a fixed value. Similar to AdaBoost, the gradient boosting algorithm's n estimator can be adjusted. The default value of n estimator for this algorithm is 100, but we can also specify the value of n estimator explicitly.

Both continuous and categorical target variables (as a Regressor) can be predicted using the gradient boosting algorithm (as a Classifier). The cost function when it is used as a classifier is Log loss, and when it is used as a Regressor, Mean Square Error (MSE) is used.

Now, using a single example, let's learn how the Gradient Boosting Algorithm functions. As shown in the example below, Likes Exercising, Goto Gym, and Drives Car are independent variables, while Age is the Target variable. Since the target variable in this instance is continuous, the Gradient Boosting Regressor is employed.

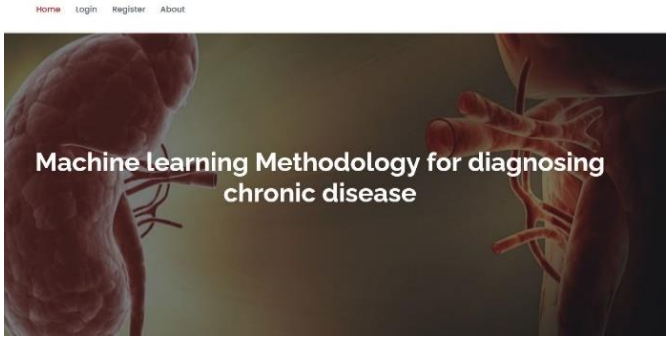
The estimator-2 will now be revealed. The residues (agei - mu) of the first estimator are treated as root nodes in the Gradient boosting algorithm, in contrast to AdaBoost, as illustrated below. Consider a scenario where a different dependent variable is predicted

using this estimator. The records with the fake Goto Gym.

V. Results and Discussion

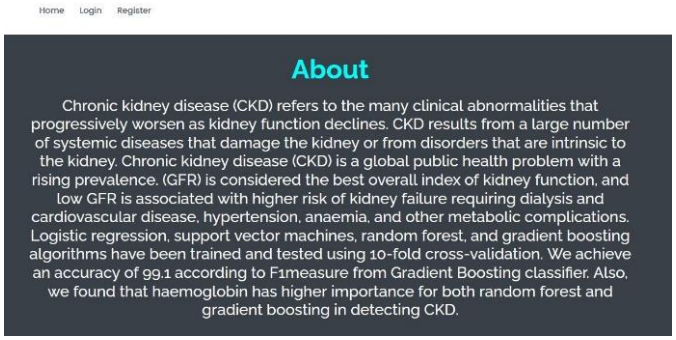
The following screenshots are depicted the flow and working process of project.

Home Page: Here user view the home page of A Machine Learning Methodology for Chronic Kidney Disease web appellation.



About page:

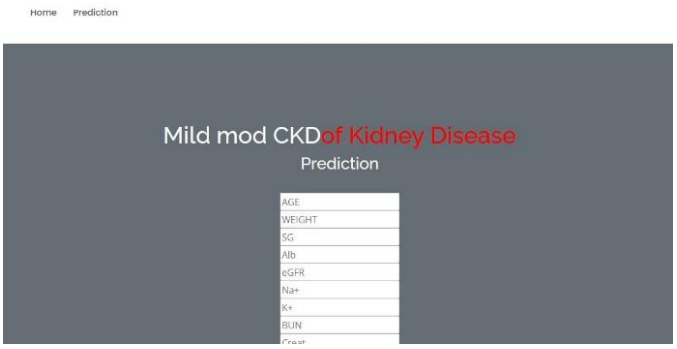
In the about page, users can learn more about A Machine Learning Methodology for Chronic Kidney Disease and symptoms of the particular disease.



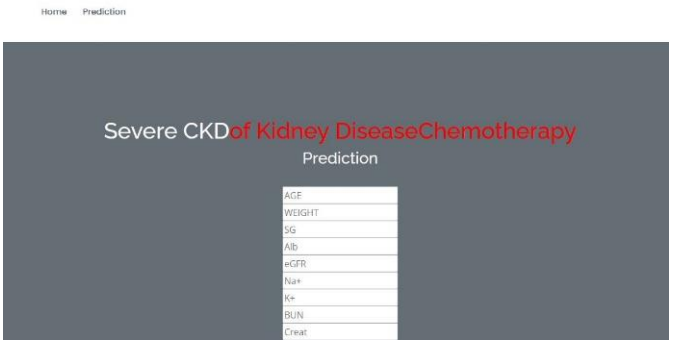
Prediction page:

This page show the detection result of The Patient has Machine Learning Methodology for Chronic Kidney Disease.

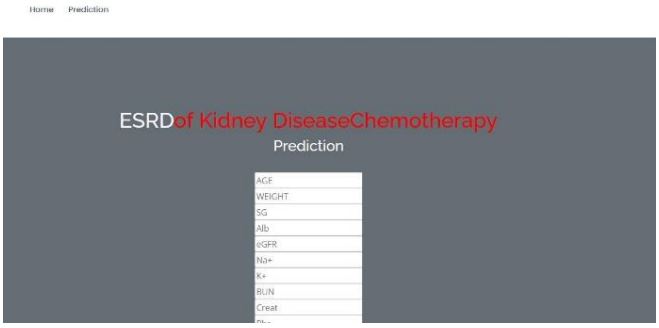
The Patient need Treatment for Mild mod CKD of kidney disease chemotherapy



The Patient need Treatment for Severe CKD of kidney disease chemotherapy



The Patient need Treatment for ESRD of kidney disease chemotherapy.



VI. Conclusion

In terms of data imputation and sample diagnosis, the suggested CKD diagnostic approach is practical. The integrated model could obtain adequate accuracy after unsupervised imputation of missing values in the data set using Gradient Boosting imputation. As a result, we believe that applying this technique to the practical diagnosis of CKD will have a positive outcome. Furthermore, this technology might be used to clinical data from various disorders in real-world medical diagnosis. However, because of the restrictions of the conditions, the available data samples are relatively small throughout the model's

development, the model's generalisation performance may be limited. Furthermore, because there are only two types of data samples in the data (ckd and notckd), set, and the model cannot diagnose the severity of CKD.

VII. REFERENCES

[1]. Z. Chen et al., "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.

[2]. L. Zhang et al., "Prevalence of chronic kidney disease in china: a cross-sectional survey," Lancet, vol. 379, pp. 815-822, Aug. 2012.

[3]. A. Singh et al., "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," J. Biomed. Inform. vol. 53, pp. 220-228, Feb. 2015.

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

[5]. C. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis," Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

[6]. V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," Am. J. Med., vol. 130, no. 12, Dec. 2017.

[7]. H. S. Xu, L.Wang, andW. L. Gan, Application of improved decision tree method based on rough set in building smart medical analysis CRM system, Int. J. Smart Home, vol. 10, no. 1, pp. 251–266, 2016.

[8]. N. R. Hill et al., "Global prevalence of chronic kidney disease – A systematic review and meta-analysis," Plos One, vol. 11, no. 7, Jul. 2016.

[9]. M. M. Hossain et al., "Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts," IEEE Trans. Ultrason. Ferr. vol. 66, no. 3, pp. 551-562, Mar. 2019.

[10]. M. Alloghani et al., "Applications of machine learning techniques for software engineering learning and early prediction of students' performance," in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246-258.

Cite this article as :
Sh