Scope
Volume 14 Number 1 March 2024

# Efficient Forecast of Chronic Kidney Disease using Gradient Boosting Classifier

**Dr. Abdul Majid[1], Shubhi Srivastava[2], DeviPrasad Mishra[3], Debashis Dev Misra[4], Veeresha R K[5], Dr G Sambasiva Rao[6].**

1. Department of Computer Science & Engineering, Yenepoya IT, Mangalore, India
2. Department of Information Science & Engineering, New Horizon, Bangalore, India
3. Department of Computer Science & Engineering, Chhatrapati Shivaji IT, Durg, India
4. Department of Computer Science & Engineering, Assam Downtown University, Guwahati, India
5. Department of Robotics & Artificial Intelligence, NMAM IT, Udupi, India.
6. Department of Artificial & Intelligence & Machine Learning, NSAK College of Engineering, Hyderabad, India

**Abstract**

Because of its fast-increasing prevalence, chronic kidney disease (CKD) is soon becoming a serious worry for the general public's health. This study's objective is to determine whether or whether machine learning methods are useful in the process of developing CRFs for chronic kidney disease (CKD) using the restricted number of clinical characteristics that are currently accessible. It has been determined via the use of several statistical procedures, including the analysis of variance, the Pearson correlation, and the Cramer's V test, that some features may be eliminated. For the purposes of training and evaluating logistic regression, support vector machines (SVMs), random forests, and gradient boosting, ten-fold cross-validation was used. When we use the Gradient Boosting classifier, we can get an accuracy of 99.1 percent using the F1-measure. In addition, we concluded that hemoglobin is a more reliable indicator of chronic kidney disease (CKD) than either random forest or gradient boosting. In conclusion, when compared to previous research, our results are among the most significant even though we have only accomplished a smaller number of characteristics so far. Because of this, the total cost of diagnosing CKD with all three tests is just $26.65. The rapidly increasing prevalence of chronic kidney disease (CKD) makes it a significant problem for the public's health. Throughout the course of this investigation, we want to test several machine learning algorithms to determine the extent to which they can diagnose chronic kidney disease based on a restricted number of clinical characteristics. Several statistical tests, including the ANOVA, the Pearson's correlation, and the Cramer's V test, have been carried out to get rid of features that aren't essential.

**Keywords:** Chronic Kidney Disease (CKD), Machine Learning (ML), support Vector Machine (SVR).

## Introduction

Chronic kidney disease (CKD) is a disorder that develops when the kidneys are damaged to the point that they are unable to filter blood effectively. This is the case when the kidneys are damaged. The primary function of the kidneys is to filter waste products and excess water from the blood that circulates throughout the body. pee is the product of this chemical process, which produces pee. If you have chronic kidney disease (CKD), it indicates that waste products have accumulated inside your body. This medical issue is chronic since it is expected to cause harm in a gradual manner over the course of time. This illness has the potential to spread to everyone, in any part of the planet. Having chronic kidney disease (CKD) raises your chance of developing a wide range of health complications. Diabetes, high blood pressure, and cardiovascular disease are the three illnesses that are responsible for most cases of chronic kidney disease (CKD). CKD is an abbreviation for chronic kidney disease. Age and gender are further factors that come into play when determining who may develop chronic kidney disease (CKD), in addition to the primary health problems listed above. You may have a broad variety of symptoms if one or both of your kidneys are not functioning properly. Some of these symptoms include back pain, stomach discomfort, diarrhoea, fever, nosebleeds, rash, and vomiting. You may also have a rash. Diabetes and high blood pressure are the two most frequent diseases that might potentially lead to kidney damage throughout the course of one's lifetime. Hence, the management of these two disorders may also be thought of as the prevention of chronic kidney disease (CKD). Many individuals who have chronic kidney disease (CKD) do not become aware that they have the condition until it is too late. This is since CKD does not often display any symptoms until it has developed into a more advanced stage.

## Stages of CKD

### Early stages of CKD

In the early stages of chronic kidney disease (CKD), patients often do not experience any symptoms. This is because the human body is capable of compensating for a significant loss of renal function, which is the reason why this is the case. It is common for kidney disease to not be discovered until this stage unless a regular test for another illness, such as a test of the blood or urine, uncovers a suspected concern. On the other hand, kidney disease may sometimes be identified at a more advanced stage. If it is detected in its early stages, therapy with medication and continuing monitoring that includes periodic testing may, with any luck, prevent it from developing into a more severe version. If it is only found later, it may be too late to stop the disease from spreading.

### CKD in its advanced stages

There may be a variety of indicators that indicate kidney disease, particularly if it is not diagnosed early or if it continues to worsen after therapy.

The last stage of chronic kidney disease is kidney failure. End-stage renal disease is another name for this condition, as is established renal failure. It is probable that at some time in the future, either dialysis or a kidney transplant will be required.

### When to see a physician

Make an appointment with your primary care physician if you notice any signs or symptoms related to renal disease. If the progression of renal disease could be diagnosed at an earlier stage, it would be possible to avoid kidney failure. If you have a medical condition that puts you at a higher risk of developing renal disease, your doctor may do blood and urine tests during office visits to evaluate your blood pressure and evaluate how well your kidneys are functioning. Inquire with your primary care physician on whether you need to undergo these tests.

### Tests for CKD

Chronic renal disease is the medical term for when an illness or condition prevents the kidneys from functioning correctly, and this may happen for several reasons. If the kidneys were functioning improperly in any manner, this is a possibility that may occur. This is a distinct possibility if the kidneys are also impacted by another illness or condition.

According to studies, the annual increase in the number of patients with chronic kidney disease who are admitted to hospitals is 6.23 percent, even though the worldwide mortality rate has remained the same. There are only a few of diagnostic procedures that may be performed to determine the stage of chronic kidney disease (CKD), and these include: I an estimated glomerular filtration rate (eGFR); (ii) a urine test; (iii) a blood pressure measurement; and (iv) testing for CKD.

### eGFR

The estimated glomerular filtration rate is a representation of the kidneys' capacity to filter blood in your body. An eGFR of more than 90 is indicative of good kidney health since it demonstrates that the kidneys are effectively filtering waste materials. If your estimated glomerular filtration rate, also known as eGFR, is less than 60, it is very probable that you have chronic kidney disease, also known as CKD.

### Urine test

In addition, the doctor will ask for a sample of your urine so that he can assess your kidney function. The kidneys are responsible for the production of urine. If your urine includes blood or protein, this is a sign that one or both of your kidneys are not working regularly. If your urine contains blood, this might indicate that you have a urinary tract infection.

### Blood pressure

Your doctor will take your blood pressure since the normal range of your blood pressure may provide valuable insight into the efficiency with which your heart pumps blood. If the patient's estimated glomerular filtration rate (eGFR) is less than 15, this indicates that they have entered the terminal stage of renal disease. Patients who are suffering from renal failure now only have two therapy options available to them: dialysis or a kidney transplant. Although none of these options is ideal, they are the only ones we have available for now. After beginning dialysis, a patient's prognosis about how long they may expect to live varies on a variety of variables, including their age, gender, the frequency and length of their dialysis treatments, the degree to which they are physically mobile, and their mental health. If it is determined that dialysis cannot be carried out effectively, the only other choice that the doctor has is to consider doing a kidney transplant. Despite this, the cost is astronomically expensive.

### Further tests

Additional testing to determine the full amount of kidney damage is not often done since it is not considered standard practices. A scan employing a computed tomography machine, an ultrasound machine, or a magnetic resonance imaging machine might be one of these options. Their mission is to examine the patient's kidneys to identify any obstructions that may be present. When examining a patient's kidneys for symptoms of illness, a tiny bit of kidney tissue is removed with a needle, and the cells in that tissue are examined under a microscope. In order to identify renal diseases, this procedure is carried out.

The use of cognitively complex systems in the realm of medicine is of the utmost significance. The vast volume of patient medical and treatment data might then be mined for hidden information via data mining, which could play a significant role in the process. This is the kind of information that medical professionals often acquire from their patients to understand more about their symptoms and devise treatment strategies that are more precise.

Objective

1. To study early-stage chronic kidney
2. To study Detection using Machine Learning

### Research Methodology

Performing Preprocessing on the Data The datasets that represent the real world today, particularly clinical datasets often have missing data, noisy data, redundant data, and inconsistent data. While working with low-quality data, one can expect to produce low-quality results. The initial phase in any machine learning application is to investigate the dataset and get familiar with its characteristics. This is done with the goal of getting the dataset ready for the modeling stage. The process that is being described here is referred to as "data preparation." 1) Extremes: Outliers are values that are significantly different from the feature's central tendency and are extreme. Errors made while entering data result in incorrect outliers, which are referred to as data noise. Since these outliers may be actual (valid) or important, medical data cannot be handled the same way as other data when dealing with outliers because these outliers cannot be ignored. Due of this, every single outlier that is discovered within the CKD dataset is investigated to establish whether it is credible. Extreme data points in this research that fall outside of the acceptable range from a medical perspective have been considered as missing data and then modified as will be detailed in the section on missing data.

These findings will be presented in the next paragraph. Box plots have been used in the CKD dataset with the goal of locating any outlying data points. As can be seen in Figure 1, there were certain blood glucose random values that were judged to be outliers since they were higher than 500 mg/dl. On the other hand, as was said before, 2008 was the year that witnessed the highest recorded level of blood glucose. The CKD data were obtained via the usage of the dataset website Kaggle. In July of 2015, the Indian Hospital was responsible for collecting this information. It is made up of a total of 400 samples, of which 250 were determined to have chronic renal illness while the remaining 150 did not. Each of these groups had their own sets of predictive characteristics that were recorded.

Examples include high blood pressure, low blood pressure, specific gravity, albumin, sugar, red blood cell count, levels of urea and creatinine, as well as levels of sodium and potassium. Additional considerations include specific gravity as well as the specific gravity of blood. Since the response data were uniformly scattered during preprocessing, it is essential that any training or test data be arbitrarily dispersed. Otherwise, it is very possible that the data would originate from the same class, which will result in inaccurate conclusions. Although there were 340 samples in the training data set, there were only 60 samples in the test data set since the 15% holdout approach was used.

SVMs, ANNs, and linear regression were the classification techniques that gave results with an acceptable level of accuracy; however, the mean Gaussian SVM produced the best results. Instead of utilizing the standard number of five, this validation was carried out using a 10-fold cross-validation approach. This was done so that efficiency might be improved. To do this, the data were first broken up into 10 separate components, each of which included 90% of the total data, and then the predicted error was compared to the error that was found. Stochastic, this occurs when two independent executions of the same machine learning algorithm each result in the acquisition of a somewhat distinct model may cause performance differences in the training iterations necessary to reach maximum accuracy.

This is because stochastic happens when the algorithm is run on the same data. In this setting, the word "stochastic" does not imply the same thing as "random." Stochastic are models that are constructed based on the available historical data and are used by machine learning algorithms.

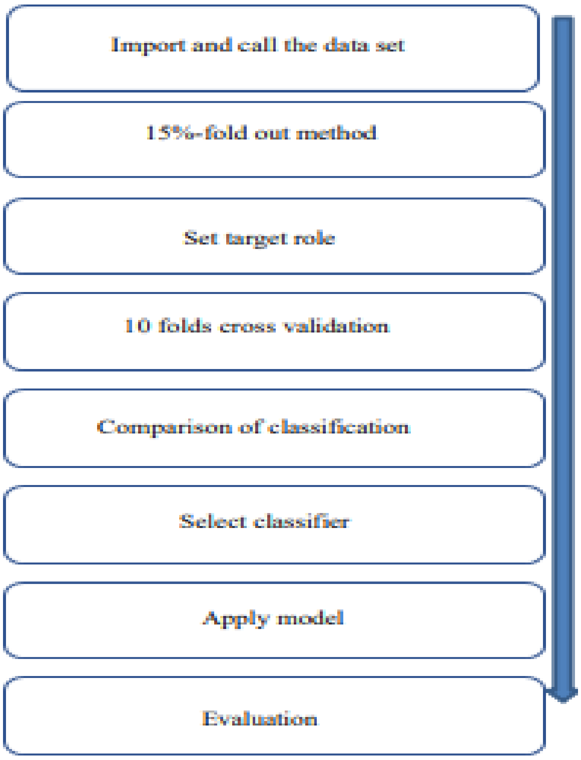

Import and call the data set
15%-fold out method
Set target role
10 folds cross validation
Comparison of classification
Select classifier
Apply model
Evaluation

**Fig.1. The whole flow processes.**

The Support Vector Machine, often known as an SVM, belongs to the subfield of machine learning known as supervised machine learning. When doing this form of training, it is vital to have a prior understanding of the category that each sample falls within. These classifiers look for hyperplanes or decision boundaries so that they can successfully separate the different classes. It accomplishes its goals via the application of the concepts of statistical learning theory and the reduction of structural risk. They are versatile enough to be used for classification as well as regression analysis. SVM works effectively in high-dimensional spaces despite its comparatively low memory requirements. The SVM technique is built on the principle of increasing the margin of error that exists between decision boundaries.

The k-nearest neighbours' method (k-NN) is an instance-based classifier that is a supervised learning approach that is comprised of the k training examples in a dataset that are physically placed the closest to one another. k-NN is also known as an instance-based classifier. The distance between a query or test point and the points in the training set that are closest to it must be calculated. This must be done before moving on to the next step. It is common practice to refer to this kind of student as a lazy learner since all the calculations are carried out within the time allotted for categorization rather than during the time allotted for learning. Using k-nearest neighbours has several advantages, the most notable of which are that it performs well even with noisy training data and does not eliminate any information in the process. The correct value of k has been determined by testing; to avoid binding, it is recommended that an odd value of k be used. k represents the number of neighbours who are the most immediate ones. Since it would make the classifier more susceptible to noise points, the value of k shouldn't be too little. On the other hand, it shouldn't be too vast since the neighbourhood may comprise points from a variety of classes. Because of this, the neighbourhood should be kept somewhat compact. The run-in which k was changed to 7 had the best performance out of all the tests that were conducted.

Using methods based on artificial neural networks, it is feasible to learn about and simulate interactions that are non-linear and complex. By carrying out many computations simultaneously, these algorithms can not only find but also define and simulate intricate connections between characteristics and classes. The neurons themselves are composed of several layers, the most notable of which are the hidden layer, which is sandwiched between the input and output layers. Most situations have several buried levels of complexity. Each feature can impact the output in several ways, such as via the application of different weights that represent the interface factors between the individual neurons; these weights need to be regularly changed in order to attain optimum performance. To make use of the ANN approach, one must first ensure that their data has been segmented and formatted appropriately. The ANN technique calls for a feature matrix to be used as input, and a response array to be used as output. This is because the ANN technique views information as a resource, which is why we get this result. The learning strategy based on trial and error had maximum training percentages of 70 percent, validation percentages of 15 percent, and testing percentages of 15 percent, respectively. Fifteen neurons were taken away in a secretion.

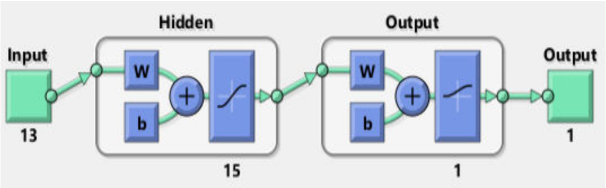


**Fig.2 Neural Network Diagram**

MATLAB 2021a was used to apply the classifiers to the same dataset, and the resulting output was a confusion matrix that included both the validation data and the test data. The results of this study are accurate and specific predictions, as well as high levels of sensitivity and precision. These equations, which can be found in numbers 1 through 4, were used to generate the following performance-evaluated parameters for the model:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

Confusion matrices are used to characterize the performance of the model. A true positive, denoted by the notation TP, is a sample of ckd that has been properly recognized. This table will be used to explain the performance of the classifier on the specified test data. A diagnosis of false negative (FN) for ckd samples suggests that the disease was not properly identified. The term "False Positive" (FP) refers to a diagnosis of the Non ckd samples that is inaccurate, while the term "True Negative" (TN) refers to a diagnostic of the Non ckd samples that is accurate. In this work, three distinct machine learning approaches are used to the problem of determining the presence of chronic renal disease (CKD). The results are summarized in TABLE I, which may be seen below. Classifiers based on artificial neural networks were able to obtain an accuracy rate of 99.2% by using effective pattern recognition and all the patient data included in the chosen data set. For the purposes of an ANN classifier exhibiting individuals with chronic renal illness who were successfully predicted as belonging to class (1). There was an error in the prediction that one patient did not have the condition even though the patient had renal failure. The forecast suggested that the patient did not have the condition (category 0).

According to the forecast, some of the patients do not have the illness, which is a correct prediction; however, the other patients were projected to have the condition, which is a wrong prediction. Similarly, there were 151 individuals who did not have the syndrome at all. When evaluating the effectiveness of the classifier, it is possible to make use of both the sensitivity and the anticipated positive value. Both the accuracy and the true positive rate reached their respective maximums of 99.2% and 99.6%, respectively. The connection between a classifier's true positive rate and its false positive rate after training may be visualized with the use of the Receiver Operating Characteristic (ROC) curve.

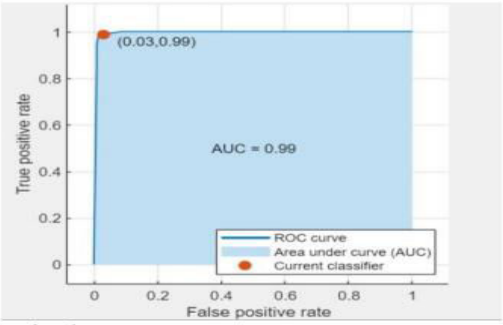
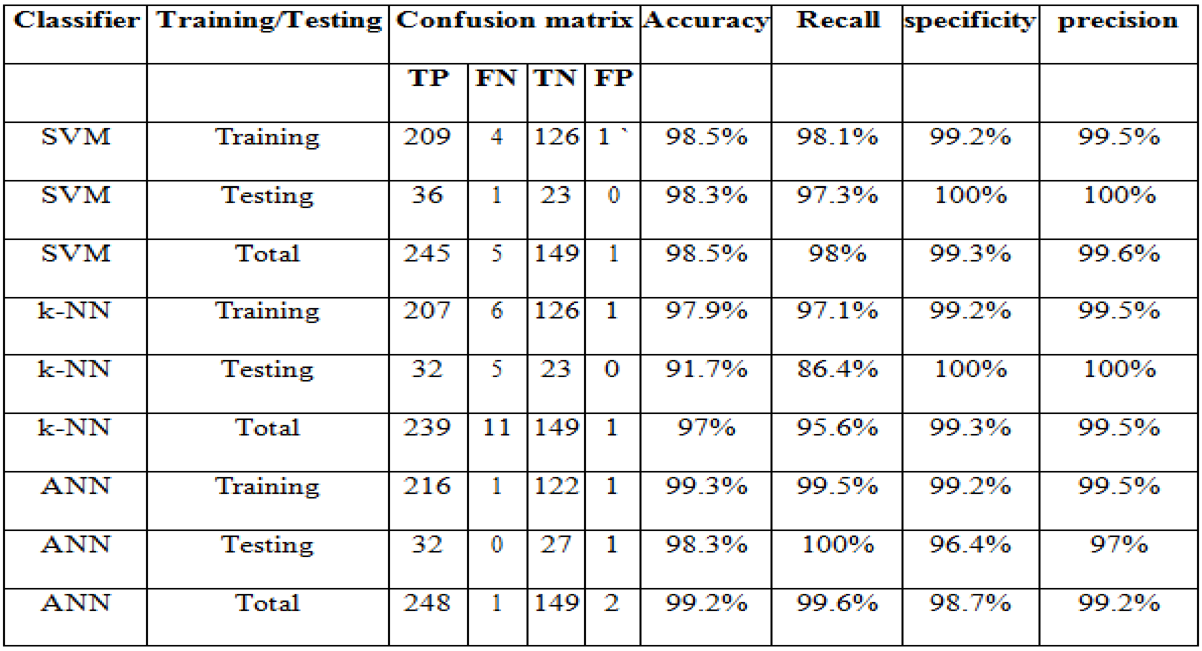


**Fig. 3 ROC for ANN classifier**

The accuracy of the SVM classifier, which was used in this investigation, was 98.5%. This is an outstanding result; nevertheless, it is not quite as good as the accuracy achieved by the ANN classifier. In accordance with the SVM, patients fell into the categories of false-negative, true-negative, and false-positive, while one forecast fell into the group of false-positive. According to the data, the accuracy is 99.6% sensitive and 98% accurate.

Table 1. Displays the results for each classifier, including training and testing outcomes, as well as the overall measured accuracy for each of themThe findings from the SVM classifier indicate that 245 samples were accurately identified as having ckd, whereas a single sample was given an inaccurate diagnosis, and 149 samples were not diagnosed at all. Accuracy Once again, we need exactness and attention to detail. There were 239 samples that had a correct diagnosis of ckd, 1 sample that had a misdiagnosis of ckd, and 149 samples that had a right diagnosis of not having ckd. However, there were 6 samples that had a

misdiagnosis of not having disease while really having disease in the confusion matrix TP FN TN FP. The ANN classifier performed far better, accurately identifying the presence of the disease in 248 of the samples. Accuracy, recall, specificity, and precision are the performance measurements that can be seen when you look at the confusion matrix. You can also observe these measures. As can be seen in the table that follows, when utilizing these metrics to assess the models, K-NN attained an accuracy of 97%, which is regarded as a good performance but is not sufficient on its own. In terms of accuracy, the SVM fared much better than the K-NN, with a rate of 98.5% overall. Nevertheless, in contrast to the K-NN and the SVM, the ANN was able to attain the highest performance (99.6%), as can be shown. The SVM classifier fared better than the K-NN classifier and the ANN classifier, both of which attained an accuracy of 99.5%. The SVM classifier had a predicted positive value of 99.6%. Because of this, the ANN classifier has the potential to attain improved performance when identifying chronic renal disease.



| Classifier | Training/Testing | Confusion matrix | | | | Accuracy | Recall | specificity | precision |
|---|---|---|---|---|---|---|---|---|---|
| | | TP | FN | TN | FP | | | | |
| SVM | Training | 209 | 4 | 126 | 1 ` | 98.5% | 98.1% | 99.2% | 99.5% |
| SVM | Testing | 36 | 1 | 23 | 0 | 98.3% | 97.3% | 100% | 100% |
| SVM | Total | 245 | 5 | 149 | 1 | 98.5% | 98% | 99.3% | 99.6% |
| k-NN | Training | 207 | 6 | 126 | 1 | 97.9% | 97.1% | 99.2% | 99.5% |
| k-NN | Testing | 32 | 5 | 23 | 0 | 91.7% | 86.4% | 100% | 100% |
| k-NN | Total | 239 | 11 | 149 | 1 | 97% | 95.6% | 99.3% | 99.5% |
| ANN | Training | 216 | 1 | 122 | 1 | 99.3% | 99.5% | 99.2% | 99.5% |
| ANN | Testing | 32 | 0 | 27 | 1 | 98.3% | 100% | 96.4% | 97% |
| ANN | Total | 248 | 1 | 149 | 2 | 99.2% | 99.6% | 98.7% | 99.2% |

Table - 1

## Conclusion

The outlook for patients with CKD is the subject of discussion in this paper. In this setting, the ACO approach is employed as a feature selection wrapper. The acronym ACO refers to an optimization technique that uses a meta-heuristic approach. Of of the 24 different qualities that are provided, only the 12 most important ones are used for the prediction. For purposes of prediction, the SVM machine learning technique is used. The results of this classification exercise are broken down using SVM into two categories: those with CKD and those who do not have it. The primary objective of this research was to reduce the number of numerical characteristics used in the prediction process while maintaining the same level of accuracy. A rate of accuracy equal to 96 percent is achieved here.

**Reference:**

1. Hussein Abbass, “Classification Rule Discovery with Ant Colony Optimization”, Research Gate Article, 2004 Mohammed Deriche, “Feature Selection using Ant Colony Optimization”, International Multi-Conference on Systems, Signals and Devices, 2009.
2. .X. Yu and T. Zhang, ”Convergence and runtime of an Ant Colony Optimization”, Information Technology Journal 8(3) ISSN 812- 5638, 2009 David Martens, Manu De Backer, Raf Haesen, “Classification with Ant Colony Optimization”, IEEE Transactions on evolutionary computation, Vol.11, No.5, 2010.
3. Vivekanand Jha, ”Chronic Kidney Disease Global Dimension and Perspectives”, Lancet, National Library of Medicine, 2013 Kai-Cheng Hu, “Multiple Pheromone table based on Ant Colony Optimization for Clustering”, Hindawi, Research article, 2015. Guneet Kaur, “Predict Chronic Kidney Disease using Data Mining in Hadoop, International Conference on Inventive Computing and Informatics, 2017.
4. SiddeshwarTekale, “Prediction of Chronic Kidney Disease Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering, 2018.
5. Baisakhi Chakraborty, “Development of Chronic Kidney Disease Prediction Using Machine Learning”, International Conference on Intelligent Data Communication Technologies, 2019.
6. J. Snegha, “Chronic Kidney Disease Prediction using Data Mining”, International Conference on Emerging Trends, 2020. Subasi, A., Alickovic, E., &Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In CMBEBIH 2017 (pp. 589- 594). Springer, Singapore.
7. Mahmood, S. W., Basheer, G. T., &Algamal, Z. Y. (2020). Classification of chronic kidney disease data via three algorithms. AlRafidain University College For Sciences, (46)