




IJCRR  
Section: Healthcare  
Sci. Journal Impact  
Factor: 6.1 (2018)  
ICV: 90.90 (2018)  
  
Copyright@IJCRR

# Prediction and Forecasting of Persistent Kidney Problems Using Machine Learning Algorithms

Debnath Bhattacharyya<sup>1</sup>, Bhanu Prakash Doppala<sup>2</sup>, N. Thirupathi Rao<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, K L Deemed to be University, KLEF, Guntur - 522502, India; <sup>2</sup>Department of Computer Science and Engineering, Vignana's Institute of Information Technology (A), Vishakhapatnam, AP, India.

## ABSTRACT

Persistent Kidney Illness is an extremely hazardous health problem that has been spreading in addition to expanding due to diversification in lifestyle such as food routines, modifications in the environment, and so on.

**Aim and Objective:** The field of health science generates substantial amounts of information from Electronic Wellness Records. According to the wellness data of India, 63538 cases have been registered on persistent kidney condition. The average age of male and female prone to renal problems occurs within the variety of Mid Forty and Seventy year age groups.

**Conclusion:** This paper's original idea is to make a comparative study on various classification techniques and their performance.

**Key Words:** Disease Forecasting, Kidney Diseases, Classification.

## INTRODUCTION

Machine learning and data processing play a vital role in getting more flexible and understandable reports on the idea of varied techniques. Kidneys role act as blood purifiers that remove waste contents while preserving new valuable blood contents like proteins. If the purifiers were damaged, the protein content would be initially leaked, and the substances may seep into urine from the blood. Sometimes the chronic renal disorder is amid high vital sign, which not only is often caused by kidney damage but also further accelerates kidney injury and maybe a significant reason for the adverse effects of chronic renal disorder on other body parts automatically increases the risk of a heart condition and heart-strokes, collection of excess body fluids, anaemia, weakening of bones and deterioration mainly the body will not support for medications. It cannot be detected until the seriousness of the disease is advanced. If detected early, treatment can hamper or refrain kidney function and deny and reduce the opposite effects on new body parts.

A biopsy measuring tool called glomerular filtration rate works on the kidneys for removing waste blood contents called creatinine. If the value lies within the range of 60 to

90, it is an early sign of occurring kidney disease; a worth below 60 is typically considered as an abnormal phase.<sup>1</sup> Testing urine samples gives the results of protein contents (albumin) within the urine; repeated results of 30 mg or more can signify a drug. Huge vital signs can also point to underlying chronic renal disorder. Distinct machine learning procedures are appropriate for analyzing the data from distinct prospects and reviewing them into useful data.

Machine Learning is an application of artificial intelligence (AI) that gives systems the capacity to use analytical strategies to give computers the ability to learn with information and improve from experience without being explicitly configured.

## Literature Survey

These days, AI calculations are generally utilized in the field of medication. Various works have been done where AI systems are utilized to predict illness (disease). Sossi Alaoui. et al. shows the utilization of AI in infection forecast over extensive information examination.<sup>2</sup> Sandeep Reddy and Jaya. et al.,<sup>3</sup> AI (ML) systems are used to research how Chronic Kidney Disease (CKD) can be analyzed. In another exploration work of Aljaaf and Ahmed J. et al.,<sup>4</sup> CKD's arrangement is finished

### Corresponding Author:

Debnath Bhattacharyya, Department of Computer Science and Engineering, K L Deemed to be University, KLEF, Guntur - 522502, India.  
Email: [debnathb@kluniversity.in](mailto:debnathb@kluniversity.in)

ISSN: 2231-2196 (Print)

ISSN: 0975-5241 (Online)

Received: 26.07.2020

Revised: 20.08.2020

Accepted: 15.09.2020

Published: 27.10.2020

utilizing Logistic Regression, Wide and Deep Learning, and Feedforward Neural Network. Different part based Extreme Learning Machines are assessed to foresee CKD in Er. Ajay Sharma and Milandeep Arora. et al.<sup>5</sup> N.Radha. et al.,<sup>6</sup> Naive Bayes, Decision Tree, K-Nearest

Neighbour and Support Vector Machine is applied to foresee CKD. Back-Propagation Neural Network, Radial Basis Function, and Random Forest are utilized to foresee Chronic Kidney Disease (CKD).<sup>7</sup> For anticipating the CKD in bolster vector machine (SVM), K-closest neighbours (KNN), choice tree classifiers,<sup>8</sup> and calculated relapse (LR) are utilized. Multiclass Decision timberland calculation performed best to foresee CKD. After utilizing Adaboost, Bagging,<sup>9</sup> and Random Subspaces group learning calculations for the finding of CKD, Chippa, M K Robinson et al.<sup>10</sup> proposed that troupe learning classifiers give better arrangement execution. Choice tree and Support Vector Machine calculations are utilized.<sup>11</sup> XGBoost based model is created for CKD forecast with better exactness,<sup>12</sup> Gera P. et al.,<sup>13</sup> J48, and arbitrary backwoods works superior to Naive Bayes (NB), insignificant successive advancement (SMO), sacking, AdaBoost calculation.

Patients with CKD are in danger of movement to *End-Stage Renal Disease* (ESRD) and expanded cardiovascular horribleness and mortality.<sup>14</sup> They worked on the way to forestalling both of these two results is an acknowledgement of the most punctual phases of kidney malady and commencement of a focused on and forceful administration plan. The National Kidney Foundation gives proof-based clinical practice rules for all phases of CKD and related complications,<sup>15</sup> which incorporate a suggestion for a referral to a nephrologist if CKD is adequately best in class. The significance of an opportune referral to a nephrologist is evident in numerous examinations that have indicated a relationship with late nephrology referral and poor results when beginning hemodialysis.<sup>16-19</sup> Patients with unrecognized CKD might be alluded by their supplier later than a patient with perceived CKD.

Just if suppliers perceive that their patients have CKD will suitable focused on the executives be started. A few agents have shown extensive under-acknowledgement by essential consideration professionals. De Lusignan and associates showed that fewer than 4% of patients with CKD had been coded as having renal disease. Studies led the manual diagram survey (bypassing the known International Classification of Diseases (ICD) - 9 coding affectability issues identified and exhibited that more than 75 % of patients with CKD were not perceived as having CKD.

An initial phase in making an instrument to incite early acknowledgment of CKD is to decide whether the supplier has perceived the patient's CKD. The instrument could scan for CKD's proper credentials in the patient's notes as an intermediary for acknowledgment. If documentation is inad-

equate with regards to, the device could incite the supplier to reconsider the patient's record along these lines, conceivably expanding familiarity with the patient's condition. Since a manual survey of notes for documentation is not doable for an enormous scope, we thought that characteristic language handling (NLP) based strategies would help determine whether patients with CKD had the determination of CKD reported in reports. A few gatherings have effectively utilized NLP strategies to discover documentation of explicit infections or conditions. We contemplated that we could utilize a similar technique to survey whether ailment credentials were available in the notes of patients with CKD.

This examination's motivation was to create, approve, and utilize a CKD-documentation-check device to decide if CKD had been fittingly recorded in special outpatient notes in the EHR.

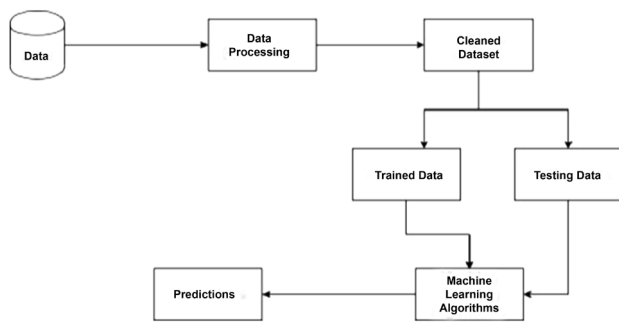
### Existing System

In the existing system, the previous predictions of the Persistent Kidney Illness are to tell the accurate values to use some of the algorithms in the previous predictions. Early acknowledgement and brief usage of prescribed administration rules are necessary to forestall intensifying kidney work and cardiovascular horribleness in patients with beginning time CKD.<sup>4</sup> A significant obstacle in accomplishing these objectives is the thing that might be the absence of acknowledgement by the essential consideration of doctors that their patients have beginning time CKD. Acknowledgement could be in a roundabout way surveyed by the nearness or non-appearance of CKD documentation comprising of words or ideas that convey the nearness of CKD. On the off chance that suppliers thinking about patients with CKD had not adequately recorded CKD in the patients' notes, at that point, the CDSS framework could advise the supplier to propose rule-based suggestions.

### Proposed System

In the suggested system, even more, forecasts must be done. Chronic Kidney Illness is a very harmful health issue that has been spreading out along with expanding because of diversification in lifestyle such as food routines, changes in the ambience, and so on. This project's main objective is to identify making use of different Classification techniques, and we need to identify the best of the classifiers as shown in Fig.1.

We select the dataset of the data containing the previous data related to a database that is used to produce accurate results or predictions that could be better than the existing system we have. So that what we have proposed in this project using the six algorithms, we will find out the accurate result better than the existing or previous.



**Figure 1:** The general execution of Machine Learning Classifier.

## Machine Learning Algorithms

### K-Nearest Neighbors Classifier Algorithm

The K-nearest neighbours are a Classification technique, and it is one of the most crucial classification techniques in machine learning. KNN belongs to the supervised learning domain and has various pattern recognition, Processing, and intrusion detection operations. Making of prognosis for a replacement datum, the data discover the nearest neighbours within the training data set. By giving the previous data, the KNN segregates the coordinates into groups classified by a selected aspect.

### Logistic Regression Algorithm

Logistic Regression is a Classification technique used for assigning observations into the discrete arrangement of classes. In general, Rectilinear Regression and Logistic Regression are very much alike. Linear Regression techniques are utilized to forecast the values, whereas Logistic Regression is employed for Classification tasks. Instead of fitting a regression curve in Logistic Regression, fitting of "S" shaped logistic function predicts two utmost values (0 or 1). Logistic Regression is often utilized for segregating the observations using various sorts of data and may easily conclude the foremost competent variables utilized for the Classification.

### Decision Tree Classifier Algorithm

Decision Tree techniques are utilized for both grouping and forecasts in AI. Utilizing the decision tree with a given arrangement of values, one can follow the different results that are an after-effect of the outcomes or choices. The decision tree is an after-effect of various steps that will assist in reaching individual choices. To assemble a decision tree, there are two stages: Induction and Pruning.

### Random Forest Algorithm

Random forest algorithm is a Classification technique, and it erects various decision trees to go about as a group of ar-

rangements and relapse process. Similarly, the random forest classifier generates a vast number of trees in the forest results in high enumerate outcomes. The main advantage of this algorithm is a reduction in over-fitting, and also, in most cases, it gives more accurate results than a decision tree. It is slow in predicting real-time data and challenging to implement.

## Support Vector Machine Algorithm

Support Vector Machine (SVM) procedure is a linear model for both the Classification and regression. SVM can be utilized to settle both linear and non-linear issues. The fundamental thought of SVM is to locate the ideal hyper-plane between the information of two classes in the preparation information.

### Stochastic Gradient Descent (SGD) Classifier

Stochastic Gradient Descent could even be a Classification machine learning algorithm that is adept for enormous large-scale learning. Stochastic Gradient Descent (SGD) is a productive methodology for linear classifiers' discriminative learning under the curved misfortune work, which is linear (SVM) and logistic regression. We apply SGD to the gigantic scope AI issues in the content arrangement and different territories of processing. It can productively scale the issue with more than  $10^5$ , preparing models furnished with more than  $10^5$  highlights. The main advantage of the SGD algorithm is very efficient and will implement these algorithms quite easily. The disadvantage is that SGD calculation requires various hyperparameters such as regularization and various cycles. SGD is also very sensitive to include scaling, one of the most significant strides under data pre-processing as shown in Fig.2.

## Algorithm

Step -1: Take the dataset that describes the data of some of the patient's health.

Step - 2: calculate the data that is compared to the gender and count of the data.

Step - 3: Take the train data to 75% and test data to 25%.

Step- 4: Training and also Examining Dataset Values utilizing different classification Algorithms.

Step - 5: Generating the Accuracy values of individual technique.

Step - 6: Comparing the performance of models.

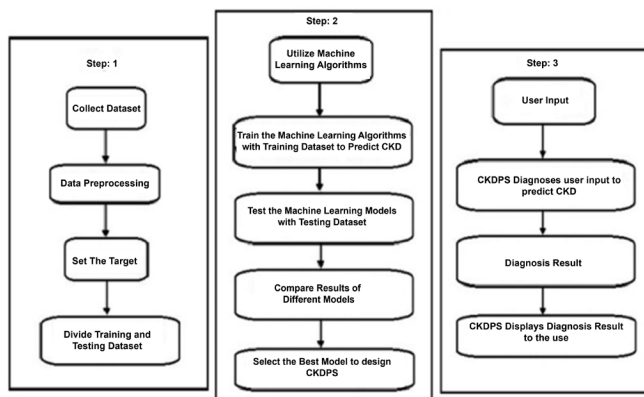


Figure 2: Execution flow of a model.

Here we take the input as the list of the data having fields related to the human body and that data are needed to be considered to the given perspective to the given data and produce the output predictions according to the data we have given below. The input will be shown in Fig. 3.

id	age	bp	sg	al	su	hemo	pcv	wc	rc	bgr	bu	sc	sod	pot	sex
1	0	48	80	1.02	1	0	15.4	44	7800	5.2	121	36	1.2	121	3.6
2	1	7	50	1.02	4	0	11.3	38	6000	3.8	151	18	0.8	142	2.7
3	2	62	80	1.01	2	3	9.6	31	7500	4.5	423	53	1.8	125	2.8
4	3	48	70	1.005	4	0	11.2	32	6700	3.9	117	56	3.8	111	2.5
5	4	51	80	1.01	2	0	11.6	35	7900	4.6	106	26	1.4	121	2.6

Figure 3: Dataset and Attributes list

## Data Set and Attributes

Experiments are directed on Chronic Kidney Disease Dataset, downloaded from the UCI Repository. This dataset contains 16 attributes (counting objective class characteristics) and 396 instances. This dataset contains information about various patients experiencing the disease. The Foremost step is data pre-processing, data transformation, and different classifiers to predict CKD and also proposes the best forecast framework for CKD. Hence to identify the best classifier, the dataset was part into two sections-Training datasets and Test dataset. Each set contains both pendants features X and output features Y. The dataset was split into 75% of training data and 25% of testing data.

## Result and Analysis

Loading the dataset into anaconda and describing the field values of the dataset is represented in Fig. 4 and 5.

```

In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

In [3]: df=pd.read_csv('kidney_disease.csv')
df.describe()

Out[3]:

```

	sg	al	su	hemo	pcv	wc	rc	bgr	bu	sc	sod	pot	sex
count	396	396	396	396	396	396	396	396	396	396	396	396	396
mean	1.028904	1.255051	0.734848	12.295707	38.257576	8133.595960	4.452273	151.000000	58.159596	3.070833	138.834596	4.554545	0.750000
std	0.066565	1.513852	1.377173	3.166023	9.321233	3284.071727	1.028720	85.431414	51.002952	5.648832	22.621174	2.877473	0.433561
min	0.000000	0.000000	0.000000	2.600000	9.000000	42.000000	1.500000	22.000000	1.500000	0.400000	4.500000	1.500000	0.000000
max	1.015000	6.000000	0.000000	10.300000	32.000000	8150.000000	3.800000	100.000000	27.000000	9.900000	133.000000	3.800000	0.750000
sum	1.020000	1.000000	0.000000	12.500000	38.000000	7800.000000	4.500000	123.500000	42.000000	1.400000	138.000000	4.300000	1.000000
var	1.025000	2.000000	1.000000	14.700000	45.000000	9600.000000	5.200000	165.000000	66.000000	2.800000	142.000000	4.900000	1.000000
std	2.825000	7.000000	8.000000	26.000000	85.000000	26400.000000	8.500000	852.000000	391.000000	76.000000	456.000000	47.000000	1.000000

Figure 4: Loading the Dataset.

```

In [4]: print(df.columns)

Index(['id', 'age', 'bp', 'sg', 'al', 'su', 'hemo', 'pcv', 'wc', 'rc', 'bgr', 'bu', 'sc', 'sod', 'pot', 'sex'],
      dtype='object')

In [5]: df.head()

Out[5]:

```

	id	age	bp	sg	al	su	hemo	pcv	wc	rc	bgr	bu	sc	sod	pot	sex
0	0	48	80	1.020	1	0	15.4	44	7800	5.2	121	36	1.2	121	3.6	1
1	1	7	50	1.020	4	0	11.3	38	6000	3.8	151	18	0.8	142	2.7	1
2	2	62	80	1.010	2	3	9.6	31	7500	4.5	423	53	1.8	125	2.8	1
3	3	48	70	1.005	4	0	11.2	32	6700	3.9	117	56	3.8	111	2.5	1
4	4	51	80	1.010	2	0	11.6	35	7300	4.6	106	26	1.4	121	2.6	0

```

In [6]: print("dimensions: {}".format(df.shape))

dimensions: (396, 16)

```

Figure 5: Data Present in the Given Dataset.

```

In [8]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396 entries, 0 to 395
Data columns (total 16 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   id           396 non-null    int64
1   age          396 non-null    int64
2   bp           396 non-null    int64
3   sg           396 non-null    float64
4   al           396 non-null    int64
5   su           396 non-null    int64
6   hemo         396 non-null    float64
7   pcv          396 non-null    int64
8   wc           396 non-null    int64
9   rc           396 non-null    float64
10  bgr          396 non-null    int64
11  bu           396 non-null    float64
12  sc           396 non-null    float64
13  sod          396 non-null    float64
14  pot          396 non-null    float64
15  sex          396 non-null    int64
dtypes: float64(7), int64(9)
memory usage: 49.6 KB

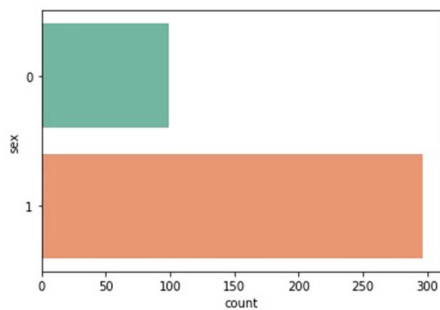
```

Figure 6: A memory that allocated to the particulars in the dataset.

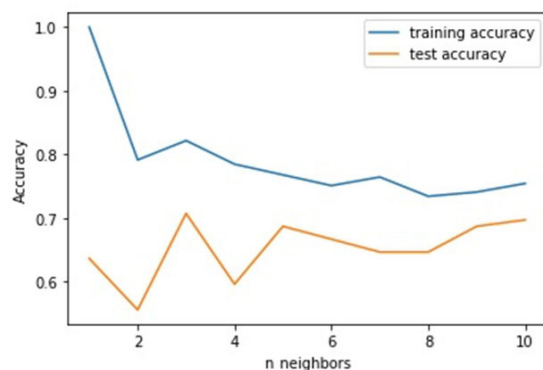
The memory type of each attribute has been represented in Figure 6. The plotting of the data based on gender had been shown in Fig. 7.



```
In [9]: sns.countplot(y=df['sex'],palette='Set2')
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x17b8156c948>
```



**Figure 7:** The graph to the plot for the given Data.



**Figure 8:** Plotting line graph for training data and test data accuracy w.r.t. n\_neighbors.

The current model's performance with both training data and the test data had been verified in terms of accuracy and had represented in figure 8 as the plot of a graphical model. The model had represented concerning the parameter n\_neighbors.

```
In [33]: from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators=100, random_state=0)
rf.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}".format(rf.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(rf.score(X_test, y_test)))

Accuracy on training set: 1.00
Accuracy on test set: 0.74

In [34]: from sklearn.svm import SVC
svc = SVC()
svc.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}".format(svc.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(svc.score(X_test, y_test)))

Accuracy on training set: 0.75
Accuracy on test set: 0.75

In [35]: from sklearn.linear_model import SGDClassifier
sgd = SGDClassifier(loss='modified_huber', shuffle=True, random_state=101)
sgd.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}".format(sgd.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(sgd.score(X_test, y_test)))

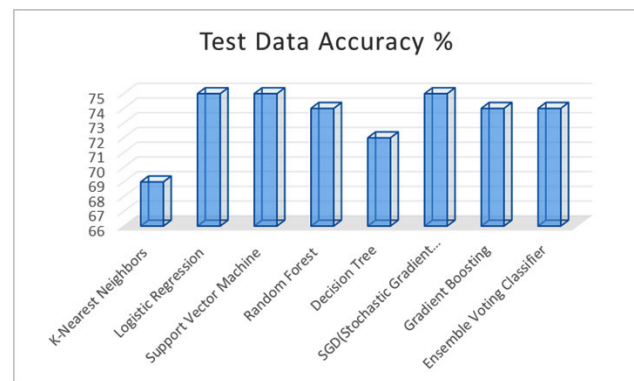
Accuracy on training set: 0.75
Accuracy on test set: 0.75
```

**Figure 9:** Accuracy on Training and Testing Data for few classifiers.

The accuracy parameter had been implemented and verified for both the training data and the test data and verified them for some of the classifiers. This [process of testing had represented in the form of pictures in Fig. 9. A comparative study on different machine learning algorithms is performed on the CKD data set. The training data exactness and testing data precision of each calculation are created to get careful and successful outcomes concerning CKD's informational index. The results of the algorithms are shown below in Table 1.

**Table 1: Comparison of values Algorithms**

Classifier	Test Data Accuracy %
K-Nearest Neighbors	69
Logistic Regression	75
Support Vector Machine	75
Random Forest	74
Decision Tree	72
SGD (Stochastic Gradient Descent) Classifier	75
Gradient Boosting	74
Ensemble Voting Classifier	74



**Figure 10:** Graphical representation of the performance of classifiers.

For a better understanding of the current model, several algorithms had been implemented with various classifiers. The performance of those classifiers had represented in the form of a graphical representation. This performance had shown in detail in the above Fig. 10. The accuracy of all these classifiers can be seen in the figure. It can be observed that the logistic regression and the support vector machine classifiers had the best results compared with the other types of classifiers among all classifiers.

## CONCLUSION AND FUTURE SCOPE

It is essential to predict Chronic Kidney Disease accurately as it is stated as a deadly disease. CKD is predicted using

six classifiers as of now. The Logistic Regression algorithm helps in the programmed location of CKD with high exactness of 75%. The exhibition of the models is assessed depending on the precision of expectations. As per the outcomes shown by all the six algorithms, the accuracy of both the training dataset and testing dataset, the Logistic algorithm gives an accurate value of 75%. As a future extension, there is a chance of applying other algorithms present in the machine learning model's classification model.

## ACKNOWLEDGMENT

Authors acknowledge the immense help received from the scholars whose articles are cited and included in references to this manuscript. The authors are also grateful to authors / editors / publishers of all those articles, journals, and books from which the literature for this article has been reviewed and discussed.

**Conflict of Interest:** Nil

**Source of Funding:** Nil

## REFERENCES

1. Radhakrishnan J, Mohan S. KI Reports and World Kidney Day. Kidney international reports. 2017 Mar 1;2(2):125-6.
2. Alaoui SS, Aksasse B, Farhaoui Y. Statistical and Predictive Analytics of Chronic Kidney Disease. In International Conference on Advanced Intelligent Systems for Sustainable Development 2018:27-38. Springer, Cham.
3. Sandeep Reddy Mula, Jaya. CKD Analysis Using Machine Learning Algorithms. International Journal for Research in Engineering Technology. 2018;6:3367-79.
4. Aljaaf AJ, Al-Jumeily D, Haglan HM, Alloghani M, Baker T, Hussain AJ, Mustafina J. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In 2018 IEEE Congress on Evolutionary Computation (CEC) 2018: 1-9. IEEE.
5. Arora M, Sharma EA. Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka. International Journal of Computer Application. 2016 Jul;6(4):20-6.
6. Classification Techniques: Hint: <https://www.edureka.co/blog/classification-algorithms/>, accessed on June 12, 2020
7. Centres for Disease Control and Prevention. Chronic Kidney Disease Surveillance System website. 2019 Last Accessed: <https://nccd.cdc.gov/CKD>. Accessed January 7, 2019.
8. National Institutes of Health. 2018 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States. Bethesda, MD: National Institutes of Health. National Institute of Diabetes and Digestive and Kidney Diseases. 2018.
9. [https://archive.ics.uci.edu/ml/datasets/chronic\\_kidney\\_disease](https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease) [Referred on 30.05.2020]
10. Chhipa MK, Robinson S, Radhouene M, Najjar M, Srimanarayana K. 2D photonic crystal microcavity ring resonator-based sensor for biomedical applications. In 2017 Conference on Lasers and Electro-Optics Pacific Rim (CLEO-PR) 2017:1-2. IEEE.
11. Jyothula, H., Rao, S. K., Vallikumari, V. Two-phase active counter mechanism embedded with particle swarm optimization technique for segmentation of biomedical images. Journal of Advanced Research in Dynamical and Control Systems (JARDCS). 2017;9(6):232-42.
12. Kishore Kumar, K., Srinath, A., Harish, M., Vijay, P., Bhaskar, K. Simulation of the four-arm parallel manipulator for medical applications. Journal of Advanced Research in Dynamical and Control Systems (JARDCS). 2017;9(18):1802-09.
13. Gera, P., Sabbisetty, V. B., Devarasetty, T., Nukala, M., Vittamsetty, N. A fuzzy preference tree-based recommender system for a medical database. International Journal of Engineering and Technology (UAE). 2018;7(1.1):319-21.
14. Nagaraju, G., Pardhasaradhi, P., Ghali, V. S. A new watermarking scheme for medical images with the patient's details. International Journal of Engineering and Technology (UAE). 2018;7(3.31):25-29.
15. Vamsidhar, E., Saichandana, B., Harikiran, J. A novel approach for feature selection and classifier optimization compressed medical retrieval using a hybrid cuckoo search. Indonesian Journal of Electrical Engineering and Informatics. 2018;6(4):410-417.
16. Aparna P, Kishore PV. Biometric-based efficient medical image watermarking in E-healthcare application. IET Image Processing. 2018 Oct 16;13(3):421-8.
17. Potharaju SP, Sreedevi M. A Novel LtR and RtL Framework for Subset Feature Selection (Reduction) for Improving the Classification Accuracy. In Progress in Advanced Computing and Intelligent Engineering 2019 (pp. 215-224). Springer, Singapore.
18. Rajendra Prasad, C., Bojja, P. A survey on routing protocols in wireless body area networks for medical applications. Journal of Advanced Research in Dynamical and Control Systems. 2018;10(10):92-97.
19. Rewar E, Singh BP, Chhipa MK, Sharma OP, Kumari M. Detection of the infected and healthy part of the leaf using image processing techniques. Journal of Advanced Research in Dynamical and Control System (JARDCS). 2017;9(1):13-9.