JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

Classification Model Based on Pathological Data for Kidney Diseases Prediction using Machine Learning Approach

S Anitha Elavarasi[1], Kannan Venkatesan[2], Murali V[3]

[1]Associate Professor, Department of Computer Science and Engineering,
Sona College of Technology, Salem District, India,
anithaelavarasi@sonatech.ac.in

[2]PG Scholar, Department of Computer Science and Engineering
Sona College of Technology, Salem District, India,
venkatkannan.90@gmail.com

[3]Senior Software Engineer
Informatica Pvt Ltd,
murali.v88@gmail.com

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

Abstract: Chronic Kidney disease (CKD) is one of major threat all over the world with high morbidity and death rate. Patients often fail to diagnose the CKD problem as it lacks the symptoms at an early stage. Early identification of CKD can help us in reducing the death rate to a high extent as well as delays the further progression of disease. Machine learning models are built to predict the presence of CKD or not. Using the pathology data on the machine-learning model helps in detecting the CKD at an early stage. In this paper different classifiers such as KNN, Naïve Bayes, Random Forest, SVM and 3DCNN algorithms are compared for its predicting accuracy of CKD. 10 cross-validation techniques are sufficient for our model random forest, an ensemble approach which combines several decision tree models and their decision are combined to make the final prediction gives a maximum accuracy and precision in predicting the chronic kidney disease

***Keywords*:** Classification, random forest, KNN, Naïve Bayes, SVM, CNN, CKD.

I. INTRODUCTION

The kidney is a vital organ in the human body whose main functions is to remove waste products from human body through the process of excretion. It not only removes the waste product it also cleans the blood, filter extra water and have a balance between water and minerals in blood. Major seven functionality [13] of kidney are (1) Balancing ACID and base level, (2) balancing water level in blood, (3) maintaining electrolyte balance in blood, (4) removing toxins from human body, (5) controlling the pressure level in blood, (6) producing the erythropoietin and (7) stimulating the proper level of vitamin D. Around 1 million cases of chronic kidney disease (CKD), also called renal failure are reported in India every year. A person can develop permanent kidney failure if we fail to diagnose CKD at an earlier stage. The patient with kidney diseases shows symptoms of weak bones, anemia, nerve damage, etc. Therefore, it is the need of the hour to detect this disease in the preliminary stages, but the symptoms are unpredictable and at times it is difficult to diagnosis in people who do not show any symptoms at all.

Data mining and machine learning techniques can be used for prediction, classification, and pattern identification applications.ML algorithms help to handle enormous amounts of data with high processing speed which makes the prediction task easy at an early stage of diseases prediction. Numerous algorithms are used in recognizing / diagnosing CKD. Machine learning algorithms can help the doctors to have early diagnosing. It uses the old CKD patient data which can be used for training the models. Some of the classification techniques employed in this paper are Logistic regression, Decision trees, Random Forest, Support vector machine, Naïve Bayes, K-Nearest Neighbor and neural network [9-13]. All classification models are trained to predict the CKD and results are compared using evaluation metrics such as accuracy, precision, recall.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

In this paper, section 2 describes the literature work related to machine learning approaches used in predicting chronic kidney diseases. Section 3 compares five different classification algorithms employed for chronic kidney diseases prediction. Section 4 describes the results of these learning models. Section 5 finally concludes the work.

II. RELEATED WORK

Ravindra et al [1] use support vector machine with radical basis kernel function based neural networks approach to classify the chronic and non-chronic kidney disease. Four different combination of feature set each consisting of three to six combination of input parameter were selected by the author for diseases prediction. The classifier was evaluated using accuracy, sensitivity, and specificity. Results shows that SVM acts as a potential candidate in classifying chronic and non-chronic kidney disease by using the six major parameters such as Albumin, Sugar, Blood glucose random, Serum Creatinine, Sodium, Potassium, Hemoglobin.

Revathy et al [2] compare the accuracy of Decision tree, support vector machine and random forest approach in predicting the chronic kidney disease. Data sets were collected from UCI machine learning repository (Chronic Disease data set) and R programming language was used for implementing all three machine learning algorithms. Results shows that random forest model achieves 99.16% accuracy and predicts CKD than the other two (decision trees and support vector machines) approaches.

Almasoud, M., & Ward, T. E[3] applies various statistical tests in identifying the most significant set of features from the data sets to be used by the machine learning algorithms for the prediction of chronic kidney disease. Statistical tests such as Pearson's correlation, ANOVA and Cramer's V were used to find the associations among various parameters. Machine learning algorithms used for comparisons are gradient boosting algorithm, SVM, Logistic regression and random forest. Among various parameter used values of hemoglobin has high impact and albumin has least impact in detecting CKD issues especially using random forest and gradient boosting algorithm.

Pasadana et al [4] compares 11 decision tree classifier algorithms for early detection of chronic kidney disease. The author uses dataset form UCI machine learning repository with 24 parameters such as age, red blood cells, blood pressure, albumin, blood glucose random, pus cell, and blood urea and so on. Data mining tool – Weka is used for implementing these entire eleven decision tree algorithm. Test were conducted using 10 fold cross validation. Results were evaluated using various parameters such as accuracy, error rate, running time, precision, recall, and kappa Statistics.

Meghanaet al [5] employs three classifier algorithms such as SVM, random forest, and neural network model for the early prediction of CKD. Data repository from Kaggle with 26 features and 400 patient data were used by the author. To create a balance in the dataset with and without CKD, input features were duplicated using a technique called synthetic minority oversampling. This approach helps in giving a better training to our selected classifier. Kunwar et al [6] employs two approaches (KNN and Ensemble SVM with Radial Basis kernel Function) for predicting CKD. The author uses the heat map, a visualization technique to find the association among various features present in the data sets. A decision tree is built by using the Gini information gain value for predicting the CKD.

Subasi et al [7] inspects the usage of machine learning algorithms for detecting the abnormalities of physiological data in diagnosing CKD. Real world data from UCI repositories is used for classification task. Results of different classifier algorithms are compared with the findings of various literatures available. Among different algorithm random forest achieves good result both in terms of quantitatively and qualitatively.

Vijayarani and Dhayanand [8] compares the performance of SVM and ANN classification algorithms in predicting four different types of kidney diseases such as Acute Nephritic Syndrome, CKD, Acute Renal Failure and Chronic Glomerulonephritis. Experiments were conducted using MATLAB and the author concludes ANN outperforms SVM.

TABLE I. LITERATURE SURVEY RELATED TO CHRONIC KIDNEY DISEASE (CKD)

Author	Machine learning Techniques	Parameters used in evaluating the classifier
Ravindra et al	SVM and neural network	Accuracy
Revathy et al	Decision tree, support vector machine and random forest	Error rate
Almasoud, M., & Ward, T. E	Statistical tests with ML	Running time

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

Pasadana et al	Decision tree algorithms	Precision
Meghana et al	SVM, random forest, and neural network model	Recall
Kunwar et al	KNN and Ensemble SVM	Sensitivity
Subasi et al	Random forest	Specificity
Vijayarani and Dhayanand& 2015	SVM and ANN	Heat map Feature selection Gini information gain for attribute selection for the decision tree node selection

III. PROPOSED MODEL

The objective of this work is to identify the patient with CKD using various machine learning classification techniques. The proposed model is shown in figure 1. Data set collections from Kaggle repository are first analyzed and preprocessing is performed to handle noise in the data. Preprocessed data is next splitted into training and testing data. ML algorithms are modeled using a python library. These models are trained and validated using the preprocessed data. Now the data from user is fed to the system and the trained model predicts whether the user is suffering from CKD or not.

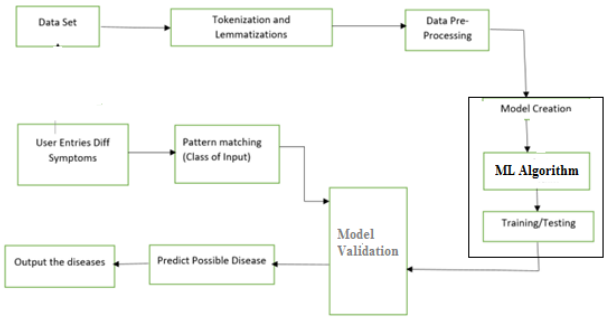

Fig 1. Chronic Kidney Diseases Prediction Model

A. Module:

1) Data Collection and Preprocessing:

Data set which was obtained from Kaggle repository. There are more than 400 records available in this database, each with 25 clinical attributes such as blood pressure, age, sugar, and many more shown in Table 2 and class label is the target class that consists of binaries values to represent the presence or absence of CKD for the patient.

TABLE II. FEATURES SELECTED FROM DATASET

S. No	Attributes	Descriptions	Type of Data
1	Age	Patients age inyears	Numerical
2	Sex	Gender of the patient	Nominal
3	Blood Pressure	Blood pressure measured in mm/Hg	Numerical
4	Red Blood Cells	Level of the RBC (normal or abnormal)	Nominal
5	Blood Glucose Random	Random Blood Glucose measured in mgs/dl	Numerical
6	Sodium	Sodium level measured in in mEq/L	Numerical
7	Potassium	Potassium level measured in in mEq/L	Numerical
8	White Blood Cell	White Blood Cell count in cells/cumm	Numerical
9	Red Blood Cell	Red Blood Cell count in cells/cumm	Numerical
10	Hypertension	Presence or absence of hypertensionfor the patient	Nominal
11	Diabetes Mellitus	Presence or absence of Diabetes Mellitus	Nominal
12	Coronary Artery Disease	Presence or absence of Coronary Artery Disease	Nominal
13	Blood Urea	Blood Urea level measured in mgs/dl	Numerical
14	Serum Creatinine	Serum Creatinine measured in mgs/dl	Numerical
15	Pus Cell clumps	Pus Cell measured in clumps	Nominal
16	Class	Classification label	Nominal

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

2) Classification:

Classification is a supervised machine learning approach used for predicting the outcome of a category based on the training labels already available for a set of categories. Different classification algorithms discussed in this paper are: KNN, RF, Decision Tree, Logistic Regression and Naive Bayes classification techniques. The dataset is splited into two phases, training, and testing. On this 80% is used for training model and 20% for testing the model. Testing dataset is used to check the performance of the trained model. Performance of each algorithm is computed and analyzed with different metrics such as accuracy, precision, recall, F1 score measures.

a) KNN Classifier

KNN classification technique simply acts as a non-parametric algorithm, i.e., it will not have any pre assumptions during data distribution analysis. During classification KNN uses an imaginary border to classify the input data points. When a new data point arrives, it attempts to find the group having 'k' closest point. KNN is often expensive in nature. Sikie-learn package is used for implementation.

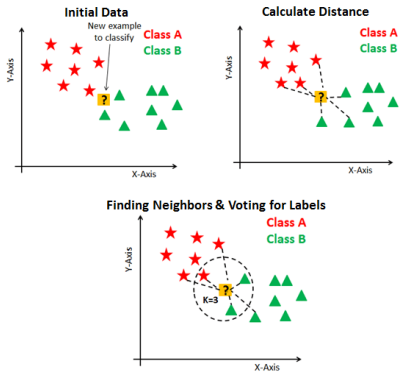

Fig 2. KNN Classifier (Google Image source [17])

Steps to be followed to implement KNN classifier algorithm:

Step 1: Import the libraries sklearn.neighbors and sklearn.metrics for k nearest neighbor and classification report respectively.

Step 2: Load the training data set

Step 3: Use the function KNeighbourClassifier and fit method to model and build the classifier. Internally inside the model the algorithm works as:

(a) Initialize any chosen number - k to represent the number of neighbors to be identified from the data

(b) For each data point, compute the distance between the test and the current input

(c) Sort the ordered collection of distances

(d) Pick the first K entries from the sorted collection

(e) Get the labels of the selected K entries and predict the label for the test data

Step 4: Validate the model by using the predict() function.

Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

b) Naïve Bayes

Naïve Bayes algorithm is based on the Bayes rule. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B) [10] as shown in equation 1:

$$P(B) = \frac{\left(P\frac{(B}{A)*P(A)}\right)}{P(B)} \qquad (1) \qquad)$$

Steps to be followed to implement Naïve Bayes algorithm:

Step 1: Import Gaussian library from sklearn .naviebayesandclassification_report library from sklearn.

Step 2: Load the training data set

Step 3: Use the function GaussianNB() to model the naïve Bayes classification approach and build the model using fit method.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

a. Internally inside the model fit it use Naïve Bayesian equation to calculate the posterior probability for each class.
b. Final predicted class will have highest posterior probability value.

Step 4: Validate the model by using the predict() function.
Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

c) Random Forest

Random Forest uses a supervised ensemble learning approach for classifying the given data points. Ensemble approach helps in improving the predictive accuracy of the learning model by combining many decisions tree model used for data classification. Random forest approach is employed to solve a complex problem as well to address the issue of overfitting. This approach can be deployed in Python and R using robust libraries.

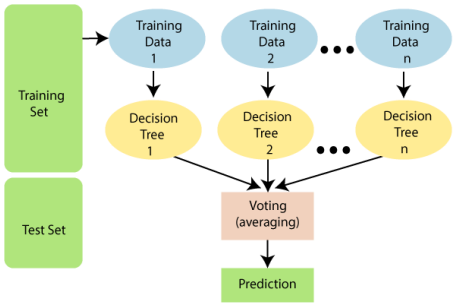

Fig 3. Random forest (Google Image source [17])

Steps to be followed to implement Random Forest algorithm:

Step 1: Import RandomForestClassifier library and classification_report library from sklearn.
Step 2: Load the dataset and start with the selection of random samples from a given dataset.
Step 3: Use the function RandomForestClassifier () and build the model using fit method. Internally the algorithm works as follow:
(a) Construct a decision tree for each sample and get a prediction result from each decision tree.
(b) Voting will be performed for every predicted result.
(c) Select the most voted prediction result as the final prediction result.
Step 4: Validate the model by using the predict() function.
Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

d) Logistic Regression

Logistic regression is supervised binary classification technique. It works for linearly separable data. i.e., if the input data is represented in two-dimensional space, we can find a straight line (Logistic regression model) separating these data into two partitions. Logistic regression suffers from overfitting if the number of observations is less than the number of features present in input data set.

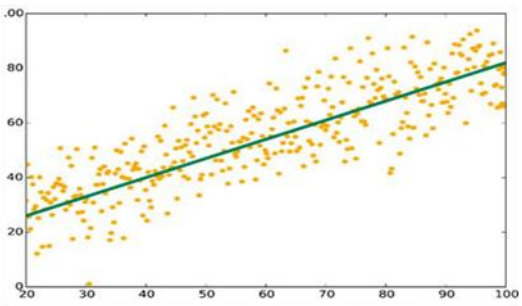

Fig 4. Regression

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

Steps to be followed to implement Logistic regression algorithm:

Step 1: Import LogesticRegression library and classification_report library from sklearn.linear model and sklearn.metrics.

Step 2: Load the training and test dataset

Step 3: Use the function LogesticRegression() and fit method to build the model. Internally the algorithm works as follow:

(a) Construct a best fitting line by identifying the relationship between independent and dependent variables.

(b) Best fit line is known as regression line and represented by a linear equation Y= a *X +b

Step 4: Validate the model by using the predict() function.

Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

e) SVM Classifier

Support Vector Machine (SVM) is a supervised machine learning model used for solving classification problems. In this approach the process of segregating the different data points into various classes is performed by identifying the correct hyper-plane. Maximizing the hyper plane margin will help to overcome the problems of misclassification. Some well-known support vector machine implementations are Sickie-learn, MATLAB, and LIBSVM.

Steps to be followed to implement SVMClassifier:

Step 1: Import the libraries sklearn.svm and sklearn. metrics for SVM classifier and classification report respectively.

Step 2: Load the training and testing data set

Step 3: Use the function svm to classify the training data and fit method to build the classifier. The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH). Internally inside the model the algorithm works with the following two steps.

a) First, SVM will generate hyperplanes iteratively that segregates the classes in best way.

b) Then, it will choose the hyperplane that separates the classes correctly.

Step 4: Validate the model by using the predict() function.

Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

f) Decision Tree Classifier

Decision tree model builds a tree like model used to predict the class of a variable by using the decision rules build using the prior data. This is a non- parametric supervised learning process. It uses basic decision rules from data features to build a model. In Decision tree each internal node represents an attribute query. Leaf node otherwise called a terminal node cannot be further splitted.

Step 1: Decide the Root node of the decision Tree by computing the entropy value of the label feature. .

Step 2: Calculate Entropy value for each of the attribute after a split.

Step 3: Computer the information gain for each split.

Step 4: Perform required number of Split.

g) 3D CNN (3D Convolution neural network)

A 3DCNN is similar to 2D CNN. The main difference exists in the working of convolution layer and pooling layer. In convolutional layer multiple pairs of 2d matrices are multiplied with different values of filters. In pooling layer instead of finding the maximum element from a 2x2 matrix, maximum element among 2x2x2 kernel is selected.

Steps to be followed to implement 3DCNN:

Step 1: Import the required sklearnlibraries for implementing 3D Convolution neural network and classification report respectively.

Step 2: Load the training and testing data set

Step 3: Use the function Con3D, maxpooling3D, Dropout and Dense function to build the model.

Step 4: Validate the model by using the fit function.

Step 5: Classification accuracy of the model is obtained by using the classification_report() function.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

IV. RESULT AND DISCUSSION

A. Data set:

Data set which was obtained from Kaggle repository [16]. There are more than 400 records available in this database, each with 25 clinical attributes such as blood pressure, age, sugar, and many more shown in Table 2 and target class that consists of binaries values to represent the presence of CKD for the patient.

B. Performance Measure / Classifier Evaluation:

Performance of the predicted model is measured using Accuracy, precision, recall and f1 score. Accuracy defines the number of correct predictions the classifier had made among the total number of test samples. Accuracy is defined using equation 2 [19]. Precision defines the number of positive predictions made. It is defined using equation 3[19]. Recall defines the number of correctly positively predicted out of all positive predictions available in the model. It is defined using equation 4[19] and f1 score.

$$\text{Accuracy} = (TP+TN)/\ (TP+FP+FN+TN) \quad (2)$$

$$\text{Precision} = (TP)/(TP+FP) \quad (3)$$

$$\text{Recall} = (TP)/(TP+FN) \quad (4)$$

Table 3 describes the accuracy of various machine learning classifier algorithms in predicting chronic kidney disease. Among the various algorithms random forest and convolution neural network achieves a maximum accuracy of 92% and 94% respectively. As random forest is a combination of multiple decision tree model and in convolution neural network a greater number of layers are employed for training and decision making, these algorithm provides more than 90% accuracy.

TABLE III. PREDICTION ACCURACY OF MACHINE LEARNING ALGORITHM

ML Algorithm	Accuracy
Decision Tree classifier	85%
Logistic Regression	84%
Random Forest	92%
Naive Bayes	82%
Support vector machine	73%
3DCNN	94%

Figure 5 describes the precision and recall of various classifier algorithms used in predicting chronic kidney disease. Among the various algorithms random forest achieves a maximum precision of 91% and all other algorithm achieves less than 85%.

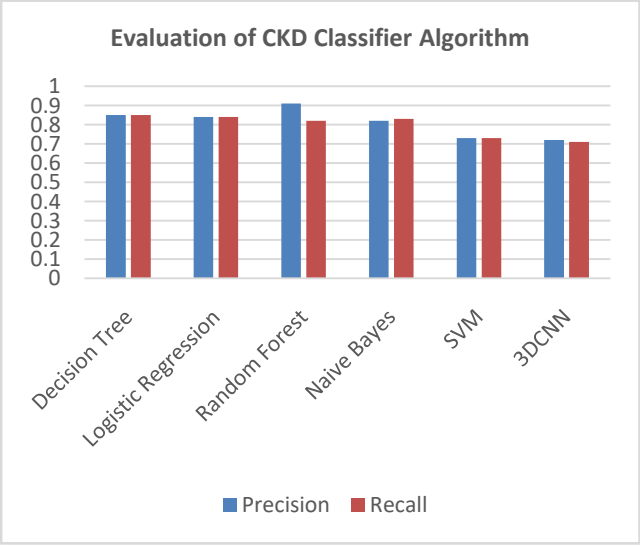

Figure 5. Precision and Recall of Various Classifier Algorithm

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 169-177
https://publishoa.com
ISSN: 1309-3452

Table 4 describes the F score of various classifier algorithms used in predicting chronic kidney disease. Among the various algorithms random forest achieves a maximum f-score of 91%, decision tree, Navie bayes and regression achieves less than 85% and SVM and neural network gives a minimal f score.

TABLE IV. F - MEASURE OF VARIOUS CLASSIFIER ALGORITHMS

Machine Learning Algorithm	F- Measure
Decision Tree	0.85
Logistic Regression	0.84
Random Forest	0.91
Naive Bayes	0.82
SVM	0.73
3DCNN	0.72

V. CONCLUSION

The Chronic Kidney Disease is undoubtedly one of the fatal diseases we are facing at present. It is particularly challenging too as its symptoms are unpredictable and at times it is difficult to diagnosis in people who do not show any symptoms at all. One of the reasons why it is hard to diagnose is that, CKD does not depend on a single pathological feature for diagnosing. Also, common symptoms of CKD are not a significant contribution in identifying the disease. In a nutshell, creating an application for detecting chronic diseases will not only help medical professionals for solving critical problems but also helps the people by reducing the screening time. Among the various predicting model employed in this paper we found that random forest, an ensemble approach which combines several decision trees models, and their decision are combined to make the final prediction gives a maximum accuracy and precision in predicting the chronic kidney disease.

REFERENCES

[1] Ravindra, B. V., Sriraam, N., &Geetha, M. J. I. J. E. T. (2018). Classification of non-chronic and chronic kidney disease using SVM neural networks. *Int. J. Eng. Technol*, *7*(1), 191-194.

[2] Revathy, S., Bharathi, B., Jeyanthi, P., & Ramesh, M. (2019). Chronic kidney disease prediction using machine learning models. International Journal of Engineering and Advanced Technology (IJEAT), 9.

[3] Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Soft Computing and Its Applications, 10(8).

[4] Pasadana, I. A., Hartama, D., Zarlis, M., Sianipar, A. S., Munandar, A., Baeha, S., &Alam, A. R. M. (2019, August). Chronic kidney disease prediction by using different decision tree techniques. In Journal of Physics: Conference Series (Vol. 1255, No. 1, p. 012024). IOP Publishing.

[5] Meghana, H. L., Kuber, V. S., Yamuna, B. S., &Varshitha, T. L. Chronic Kidney Disease Prediction using Neural Network and ML Models,International Journal of Engineering Research & Technology (IJERT),ISSN: 2278-0181,2021.

[6] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016, January). Chronic Kidney Disease analysis using data mining classification techniques. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 300-305). IEEE.

[7] Subasi, A., Alickovic, E., &Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In CMBEBIH 2017 (pp. 589-594). Springer, Singapore.

[8] Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2), 1-12.

[9] Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Ahmed, S. T., Devi, B. A., &Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. Measurement, 171, 108838.

[10] Basker, N., Theetchenya, S., Vidyabharathi, D., Dhaynithi, J., Mohanraj, G., Marimuthu, M., & Vidhya, G. (2021). Breast Cancer Detection Using Machine Learning Algorithms. Annals of the Romanian Society for Cell Biology, 2551-2562.

[11] Marimuthu, M., Vidhya, G., Dhaynithi, J., Mohanraj, G., Basker, N., Theetchenya, S., &Vidyabharathi, D. (2021). Detection of Parkinson's disease using Machine Learning Approach. Annals of the Romanian Society for Cell Biology, 2544-2550.

[12] S. AnithaElavarasi, J. Jayanthi, N. Basker, T. Jayasankar," Methods for Improving the Predictive Accuracy of Autism Spectrum Disorder Screening using Machine Learning Algorithms", International Journal of Advanced Science and Technology , ISSN: 2005-4238 , Vol. 29, No. 03, 2020, pp. 9255-9262.

[13] A.Sheryl Oliver, T.Jayasankar**,** K.R.Sekar, T.Kalavathi Devi, R. Shalini, S. Poojalaxmi and N.G.Viswesh, "Early Detection of Lung Carcinoma Using Machine Learning", Intelligent Automation & Soft Computing, Vol.30, no.3, 2021, pp.755-770

[14] https://flkidney.com/the-7-functions-of-the-kidneys/

[15] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[16] https://www.kaggle.com/mansoordaku/ckdisease?select=kidney_disease.csv.

[17] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[18] https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.javatpoint.com%2Fmachine-learning-random-forest-algorithm&psig=AOvVaw31SN-DITIL8egPb7cKHvqH&ust=1644153331963000&source=images&cd=vfe&ved=0CAkQjhxqFwoTCIi5_s7S6PUCFQAAAAAdAAAAABAD

[19] D. Venugopal, T. Jayasankar, N. Krishnaraj, S. Venkatraman, N. B. Prakash and G. R. Hemalakshmi, "Multifactorial Disease Detection Using Regressive Multi-Array Deep Neural Classifier", *Intelligent Automation & Soft Computing(2021)*, Vol.28,no.1,2021,pp.27-38.