Random Forest Based Chronic Kidney Disease Prediction and Analysis

[1]Nuvvula Nikhitha, [2]K.Durga Devi
[1,2]Dept. of CSE, KIET, Kakinada, AP, India

Abstract

Chronic Kidney Disease (CKD) need to be diagnosed earlier before kidneys fail to work.In order to help doctors or medical experts in prediction of CKD among patients easily, this paper has developed an intelligent system named Chronic Kidney Disease Prediction System (CKDPS) that can predict CKD among patients. The proposed system predict the CKD with minimal feature input instead of dumping all the features which may not relevant to predict the disease.To achieve this we have planned to approach by three feature selection algorithm with combination of two feature Extraction algorithm.After performing feature selection and Feature Extraction, those features will be trained with different Machine Learning algorithm. The accuracy of best combination algorithm will be implemented for predicting the CKD.Finally, Random Forest algorithm is chosen to implement CKDPS as it gives 95% accuracy, precision and recall results.

Keywords

Chronic Kidney Disease (CKD)· Chronic Kidney Disease Prediction System (CKDPS) · Machine learning algorithms · Random Forest Algorithm · User input · System Output.

I. Introduction

In today's world Chronic Kidney Disease (CKD) becomes one of the most serious public health problems. CKD is the damaging condition of kidneys function that can be worse over time.If the kidneys are damaged very badly then they fail to work. This is known as kidney failure, or end-stage renal disease (ESRD). The presence of CKD is found by using albuminuria test, Imaging test or Glomerular Filtration Rate (GFR) test As per report from [1], in India 6000 renal transplants are done annually. It is also reported that, CKD among people is increasing rapidly all over the world.Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Your kidneys filter wastes and excess fluids from your blood, which are then excreted in your urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in your body. In the early stages of chronic kidney disease, you may have few signs or symptoms. Chronic kidney disease may not become apparent until your kidney function is significantly impaired. Treatment for chronic kidney disease focuses on slowing the progression of the kidney damage, usually by controlling the underlying cause. Chronic kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant.We can predict the early stage of Chronic kidney Disease using machine learning techniques.The machine learning algorithms such as, kNearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) algorithm are applied on the dataset. Experimental result shows that only Random Forest Classifier algorithm gives better classification performance with 95% accuracy, precision and recall results. So, here, only Random Forest Classifierhas been used to predict CKD in CKDPS. As a screening tool the graphical user interface(GUI) of CKDPS has been developed with the help of python so that doctors or medical experts diagnose CKD among patients very easily.The rest of the paper is arranged as follows: Sect. 2 focuses on previous related works, Sect. 3 on methodology and system implementation Sect. 4 discusses about the proposed model of CKDPS, Sect. 5 shows accuracy comparison between previous related papers and the experimental results of this paper, Sect. 6 depicts the simulation result, and at last Sect. 7 ends up with conclusion.

II. Previous Related Works

Nowadays, machine learning algorithms are widely used in the field of medicine. Numerous works have been done where machine learning techniques are used to predict disease. The paper [2] shows the usage of machine learning in disease pre- diction over big data analysis. In the paper [3], machine learning (ML) techniques are used to investigate how CKD can be diagnosed. In another research work [4], classification of CKD is done using Logistic Regression, Wide & Deep Learning and Feed forward Neural Network. Various kernelbased Extreme Learning Machines are evaluated to predict CKD in [5]. In [6], Naive Bayes, Decision Tree, K-Nearest Neighbour and Support Vector. Machine are applied to predict CKD. In the paper [7], Back-Propagation Neural Network, Radial Basis Function and Random Forest are used to predict CKD. For predicting the CKD in [8] support vector machine (SVM), K-nearest neighbors (KNN), decision tree classifiers and logistic regression (LR) are used. Multiclass Decision forest algorithm performed best in [9] to predict CKD. After using Adaboost, Bagging and Random Subspaces ensemble learning algorithms for the diagnosis of CKD, the paper [10] suggests that ensemble learning classifiers provide better classification performance. Decision tree and Support Vector Machine algorithm are used in [11]. XGBoost based model is developed in [12] for CKD prediction with better accuracy. In [13], J48 and random forest works better Processing Detail in Step 1

- Collect Dataset: The dataset is collected which is obtained by the survey of CKD in India that contains laboratory results of both positive and negative cases of CKD. It contains cases of 400 patients with 25 attributes (eg, red blood cell count, white blood cell count, etc.), detail in Table 1. Input attributes are used to take input from user in CKDPS and output attribute is for diagnosis result.
- Data Preprocessing: CKD dataset contains some attributes with Nominal data, some attributes with Numerical data and some attribute with Null values. Data pre- processing is a data mining technique that is used to transform incomplete raw data in a useful and efficient format. Three steps in Data Preprocessing are shown below:
- Data Transformation: All of the nominal data are converted into numerical data. For example, Red Blood Cells & Pus Cell values: Normal = 1; Abnormal = 0. Pus Cell Clumps & Bacteria Values: Present = 1, Not present = 0.
- Missing data handle: The null values of an attribute are replaced with the calculated mean value of the attribute.

This strategy is applied on each attributes that contain null value.

- Rearrange the dataset: Each patient's records are repositioned haphazardly.

Set the Target: To classify that patient has CKD or Not CKD according to the values of input attributes the output attribute "Classification" is set than Naive Bayes (NB), minimal sequential optimization (SMO), bagging, AdaBoost algorithm.

III. Methodology and System Implementation

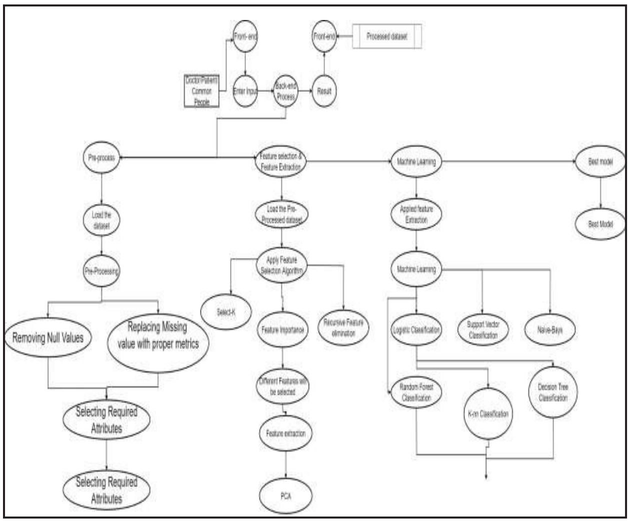

Front-end
Front-end
Processed dataset
Doctor/Patient Common People
Enter Input
Back-end Process
Result
Pre-process
Feature selection & Feature Extraction
Machine Learning
Best model
Load the dataset
Load the Pre-Processed dataset
Applied feature Extraction
Best Model
Pre-Processing
Apply Feature Selection Algorithm
Machine Learning
Select-K
Feature Importance
Recursive Feature elimination
Logistic Classification
Support Vector Classification
Naive-Bayes
Removing Null Values
Replacing Missing value with proper metrics
Different Features will be selected
Random Forest Classification
K-nn Classification
Decision Tree Classification
Selecting Required Attributes
Feature extraction
Selecting Required Attributes
PCA

Table 1. Dataset-attributes with their values.

Attribute	Abbr	Values
Age	age	2–90
Bloodpressure (mm/Hg)	bp	50–100
Specific Gravity	sg	1.005–1.025
Albumin	al	0–4
Sugar Degree	su	0–4
Red Blood Cells	rbc	Normal, Abnormal
Pus Cell	pc	Normal, Abnormal
Pus Cell Clumps	pcc	Present, Not present
Bacteria	ba	Present, Not present
Blood Glucose Random (mgs/dl)	bgr	22–490
Blood Urea (mgs/dl)	bu	1.5–391
SerumCreatinine (mgs/dl)	sc	0.4–7.6
Packed Cell Volume	pcv	0–54
Potassium (mEq/L)	pot	0–4.7
Sodium (mEq/L)	sod	0–163
Hemoglobin (gms)	hemo	0–17.8
White Blood Count (cells/cumm)	wbcc	0–26400
Red blood cell count (millions/cmm)	rbcc	0–8
Hypertension	htn	Yes, No
Diabetes Mellitus	dm	Yes, No
Coronary Artery Disease	cad	Yes, No
Appetite	appet	Good, Poor
Pedal edema	pe	Yes, No
Anemia	ane	Yes, No

Table 2. Dataset-division

Dataset	Total records	patient's	Class	
			CKD	Not CKD
Training set	320		183	137
Testing set	80		66	14

Processing Detail in Step 2

A. Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is difficult to analyze every attribute so feature selection select the most important attribute for prediction. It also increases the performance of model. In feature selection we use Select -KBest algorithm, it was simply retains the first K features of X with highest scopes. And also we used RFE (Recursive Feature Elimination) algorithm to remove the less importance features. RFE plays a main role in prediction model.

B. Feature Extraction

Feature extraction is the process of reduce the number of features and also it create a new feature from the existing one. It also extracts the new patterns available in the features. In feature extraction we use PCA (Principal Component Analysis), it combines the same attributes and also create a new patterns which is superior to original attributes. It does not combine the attribute it just evaluates the quality, predictive power and selects the best one for model.

C. Machine Learning

In machine learning algorithms we use several algorithms like Random Forest, Decision Tree, knn, SVM and Logistic Classification. Applied all these algorithms and find the accuracy of the model. Random Forest algorithm are learning method for classification, regression and other tasks that operate by constructing a decision tree at training time and outputting the class which the model of the classes or mean. In this algorithm it randomly select K features from total n features and also calculate the decision node and daughter node using best split, this process is looped until it reach the result.

Decision tree is a flowchart structure in that node represents the test on a attribute and each branch represent the outcome of the test. It only contains conditional and control statements. This algorithm split the datasets into smaller subset at the same time a associated decision tree incrementally developed. Next algorithm is K-nearest algorithm; t is a non-parametric method for classification and regression. This algorithm stores all available cases and also classifies new cases on similarity measures. Support Vector Machine is a supervised learning model with associated learning algorithms that analyze the data. It is simply the co-ordinates of individual observations. Logistic regression is a predictive analysis algorithm and based on the concept of probability.

IV. Proposed System

This section provides detail about the model of Chronic Kidney Disease Prediction System (CKDPS). It is an inquiryresponse model. Figure 2 shows the architectural components of CKDPS model. Three main modules of this model are, Administrator Module, Computational Module, and Data Processing Module. CKD Dataset contains 400 samples of the disease diagnosis. User and Administrator are two components of the model, who have interaction with the system through system's User Interface. The work of individual components is given below:

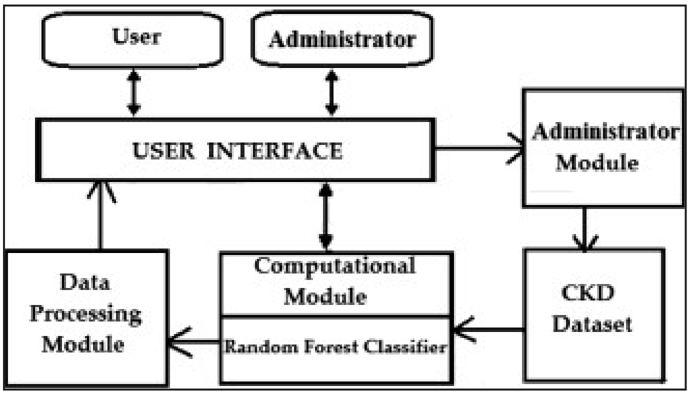


Fig. 2: Architectural components of CKDPS model

1. User Interface

User Interface is responsible to make a connection between user and the system. Through the User Interface, CKDPS displays a query form to the user and user fills up the form with values and submits it to the system. 2. User: User may be doctors or the medical experts. They provide patient's age, urine and blood test related data to the system through the User Interface. 3. Computational Module: Using Random Forest model Computational Module classifies whether the newly posted inputted data is in CKD class or Not CKD class. And it also calculates the system accuracy to perform the diagnosis. 4. Data Processing Module: After the complete classification performed by the Computational Module, Data Processing Module checks the diagnosis result. If it finds CKD class then it shows a message to the user that "Patient is suffering from CKD" otherwise it shows "Patient is not suffering from CKD". This module also displays the system accuracy to the user. 5. Administrator Module: Administrator Module assists the Administrators for administering CKDPS. Only Administrators have the permission to add, delete, update and modify the CKD Dataset records. 6. Administrator: Administrator should be doctors or medical experts, who should have proper knowledge about CKD. They can update the Dataset with valid data or delete unnecessary data from Dataset.

5. Accuracy Comparison and Experimental Results

Before the implementation of CKDPS, different machine learning algorithms are applied on the dataset and their performances are compared to the matter of accuracy, precision and recall results. Table 1 shows different accuracy, precision, recall results obtained from different machine learning algorithms. This experiment also shows Random Forest Classifier algorithm gives better performance.

Random Forest with SelectKFeature gives good result optimized model, others or overfitting.RFE feature selection selected the binary data type and also overfitting have rasied.

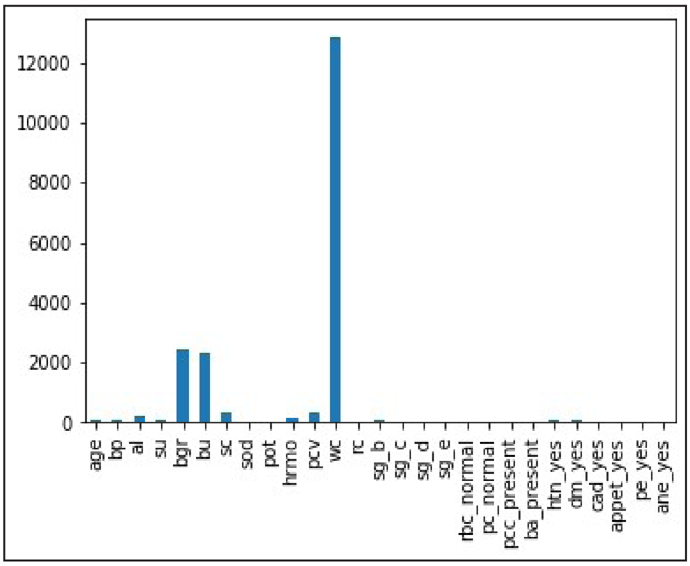
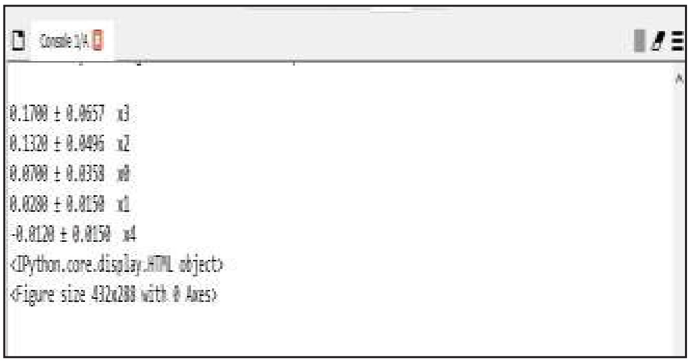


Fig. Select K Feature

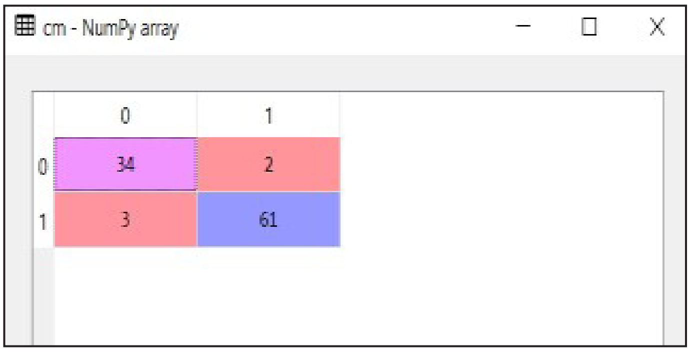
Table 3

Algorithm	Selectk	Rfe	Select_PCA	Rfe_PCA
SVM	0.95	0.98	0.82	0.98
Random Forest	0.95	0.98	0.94	0.98
Decision Tree	0.94	0.98	0.82	0.94
Navies Bay	0.98	0.98	0.98	0.98
Knn	0.89	0.98	0.81	0.98



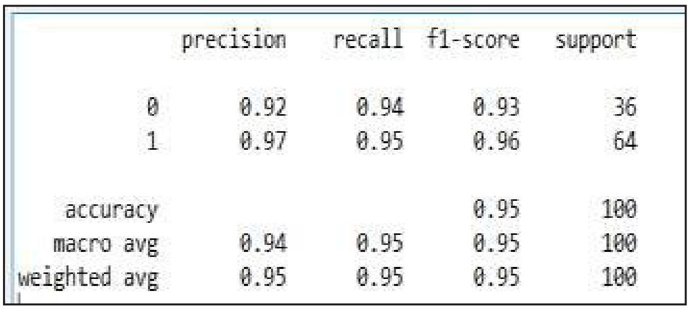
```
0.1700 ± 0.0657  x3
0.1320 ± 0.0496  x2
0.0700 ± 0.0358  x0
0.0280 ± 0.0150  x1
-0.0120 ± 0.0150  x4
<IPython.core.display.HTML object>
<Figure size 432x288 with 0 Axes>
```

Fig. Feature Weight for Select K



	0	1
0	34	2
1	3	61

Fig. Confusion Matrix



```
              precision    recall  f1-score   support

           0       0.92      0.94      0.93        36
           1       0.97      0.95      0.96        64

    accuracy                           0.95       100
   macro avg       0.94      0.95      0.95       100
weighted avg       0.95      0.95      0.95       100
```

Fig. Classification Report

VI. Simulation Result

As the simulation tool, in this paper Spyder Notebook is used in the Python Environment. To create the GUI, Tkinter method in Python has been used. GUI of CKDPS is presented in the Figs. 5 and 6.

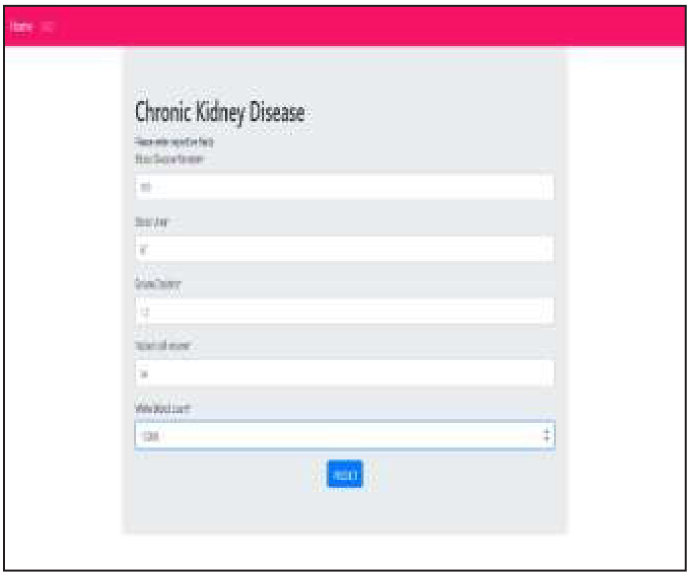


Fig. 5. According to user input patient has CKD

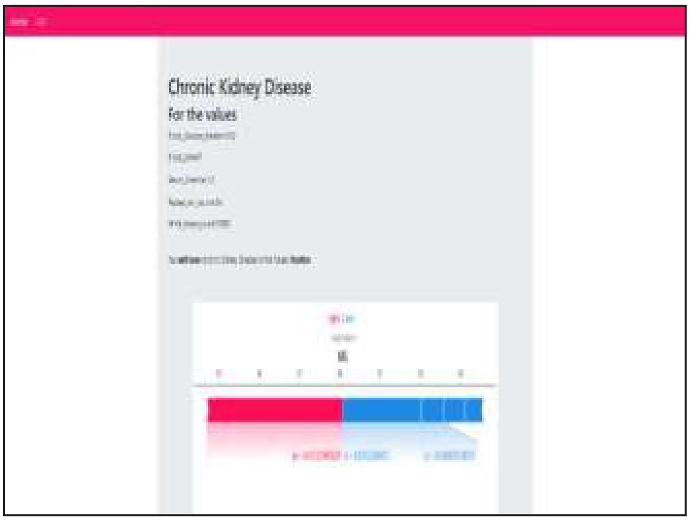


Fig. 6. According to user input patient has Positive in CKD

VII. Conclusion

Chronic Kidney Disease (CKD) should be diagnosed earlier before kidneys fail to work. To help doctors or medical experts in prediction of CKD among patients easily, this paper has developed a Chronic Kidney Disease Prediction System (CKDPS) using Random Forest Algorithm. Random Forest Algorithm is a machine learning algorithm that combines decision trees to get more accurate and stable prediction. The dataset contains cases of 400 patients with 25 attributes (e.g., red blood cell count, white blood cell count, etc.). The method of system implementation follows three steps; the primary step to implement CKDPS is Collection of Dataset, Preprocessing of Dataset, Setting of

Target class and division of Dataset into Training and Testing Datasets. In the second step machine learning algorithms such as, k-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), MultiLayer Perceptron (MLP) algorithm are applied on the CKD Dataset. Each of the different machine learning algorithms provides different accuracy, precision and recall results. From which Random Forest algorithm is selected to develop CKDPS as it provides 95% accuracy, precision and recall results. At the last step, user posts query to the system and system feedbacks the user with classification result. There are seven architectural components in CKDPS model such as User Interface, Administrator Module, Computational Module, Data Processing Module, User, Administrator and CKD Dataset. In this paper, accuracy results from the previous related works are compared. Comparison shows Random Forest algorithm works better to predict CKD.

References

[1] Agarwal, S.K.: Chronic kidney disease in India - magnitude and issues involved (2009)

[2] Vinitha, S., Sweetlin, S., Vinusha, H., Sajini, S.: Disease prediction using machine learning over big data. Comput. Sci. Eng. Int. J. (CSEIJ) 8, 1–8 (2018).

[3] Subas, A., Alickovic, E., Kevric, J.: Diagnosis of chronic kidney disease by using random forest, pp. 589–594 (2017).

[4] Imran, A.A., Amin, M.N., Johora, F.T.: Classification of chronic kidney disease using logistic regression, feedforward neural network and wide and deep learning. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–6 (2018).

[5] Radha, N., Ramya, S.: Performance analysis of machine learning algorithms for predicting chronic kidney disease. Int. J. Comput. Sci. Eng. Open Access 3, 72–76 (2015).

[6] Ramya, S., Radha, N.: Diagnosis of chronic kidney disease using machine learning algorithms. Int. J. Innov. Res. Comput. Commun. Eng. 4, 812–820 (2016).

[7] Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., Ninchawee, N.: Predictive analytics for chronic kidney disease using machine learning.

[8] Gunarathne, W.H.S.D., Perera, K.D.M.,

[9] Kahandawaarachchi, K.A.D.C.P.: Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 291–296 (2017).

[10] Basar, M.D., Akan, A.: Detection of chronic kidney disease by using ensemble classifiers. In: 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), pp. 544–547 (2017).

[11] Tekale, S., Shingavi, P., Wandhekar, S., Chatorikar, A.: Prediction of chronic kidney disease using machine learning algorithm. Int. J. Adv. Res. Comput. Commun. Eng. 7, 92–96 (2018).

[12] Ogunleye, A., Wang, Q.: Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease. In: 2018 IEEE 14th International Conference on Control and Automation (ICCA), pp. 805–810 (2018).

[13] Sisodia, D.S., Verma, A.: Prediction performance of individual and ensemble learners for chronic kidney disease. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 1027–1031 (2017).

[14] Jadhav, S.D., Channe, H.P.: Comparative study of K-NN, Naive Bayes and decision tree classification techniques. Int. J. Sci. Res. (IJSR) 5, 1842–1845.