Chronic Kidney Disease Prediction using Machine Learning Models

S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh

Abstract*: *The field of biosciences have advanced to a larger extent and have generated large amounts of information from Electronic Health Records. This have given rise to the acute need of knowledge generation from this enormous amount of data. Data mining methods and machine learning play a major role in this aspect of biosciences. Chronic Kidney Disease(CKD) is a condition in which the kidneys are damaged and cannot filter blood as they always do. A family history of kidney diseases or failure, high blood pressure, type 2 diabetes may lead to CKD. This is a lasting damage to the kidney and chances of getting worser by time is high. The very common complications that results due to a kidney failure are heart diseases, anemia, bone diseases, high potasium and calcium. The worst case situation leads to complete kidney failure and necessitates kidney transplant to live. An early detection of CKD can improve the quality of life to a greater extent. This calls for good prediction algorithm to predict CKD at an earlier stage . Literature shows a wide range of machine learning algorithms employed for the prediction of CKD. This paper uses data preprocessing,data transformation and various classifiers to predict CKD and also proposes best Prediction framework for CKD. The results of the framework show promising results of better prediction at an early stage of CKD

Keywords*: *Chronic Kidney Disease, Decision Tree, Machine Learning, Random Forest, Support Vactor s.

I. INTRODUCTION

The disability of the kidneys to perform their regular blood filtering function and others is called Chronic Kidney Disease (CKD). The term "chronic" describes the slow degradation of the kidney cells over a long period of time. This disease is a major kidney failure where the kidney sans blood filtering process and there is a heavy fluid buildup in the body. This leads to alarming increase of potasium and calcium salts in the body. Existence of high levels of these salts result in various other ailments in the body.

The prime job of kidneys is to filter extra water and wastes from blood. The efficient functioning of this process is important to balance the salts and minerals present in our body. The righ balance of salts are necessary to control blood pressure, activate hormones, build red blood cells, etc. A high concentration of calcium leads to various bone diseases and cystic ovaries in women. CKD also may lead to sudden illness or alergy to certain medicines. This state is called as Acute Kidney Injury(AKI). An increased blood pressure may lead to heart problems and heart attacks.

CKD in many cases leads to permenent dialysis or kidney transplants. A history of kidney disease in the family also leads to high probability of CKD. Literature shows that almost one out of three people diagnosed with diabetes have CKD. Literature also presents evidences of early identification and care of CKD can improve the quality of the patients life. Prediction algorithms in machine learning can be intelligently used to predict the occurence of CKD and presents a method of early medication. The detailed review on literature shows the application of various machine learning algorithms to predict CKD. This paper tries to predict CKD using the classifiers like Decision Tree,Random Forest and Support Vector Machine and also suggested best prediction model

II. LITERATURE SURVEY

M. P. N. M. Wickramasinghe.et al[1] presents a methodology to control the disease using a suitable diet plan.In this research classifiers are constructed using different algorithms like Multiclass Decision Jungle, Multiclass Decision Forest, Multiclass Neural Network and Multiclass Logistic Regression. An allowable potassium zone is predicted depending on the blood potassium levels of the patient. The classification algorithms recommend a diet place based on the predicted potasium zone.

H. A. Wibawa.et al[2] proposed and evaluated Kernel-based Extreme Learning Machine(ELM) to predict Chronic Kidney Disease. Performance of four kernels-based ELM, namely RBF-ELM, Linear-ELM, Polynomial-ELM, Wavelet-ELM are compared with the performance of standard ELM. The above methodologies were compared on metrics of sensitivity and specificity. Radial Basis Function – Extreme Learning Machine (RBF-ELM) showed higher prediction rates.

CKD increases the risk factors of CardioVascular Disease (CVD) like hypertension, diabetes mellitus, dyslipidemia, and metabolic syndrome. CKD also leads to End Stage Renal Disease (ESRD) which has no cure. U. N. Dulhare.et al[3] extracted action rules based on stages but also predicted CKD by using naïve bayes with OneR attribute selector which helps to prevent the advancing of chronic renal disease to further stages.

It is said that the median survival time of past due-stage patients is simplest approximately three years. Evaluating exactly the condition of sufferers is of incredible importance as it might substantially assist to decide appropriate care, medications or medical interventions wished, which amongst them have a complicated interrelationship and have an impact on the final results of the

Revised Manuscript Received on October 15, 2019

***** Correspondence Author

S,Revathy*, Information Technology, Sathyabama Institute of Science and Technology, Chennai, India. Email:revathy.it@sathyabama.ac.in

B.Bharathi, Information Technology, Sathyabama Institute of Science and Technology, Chennai, India. Email:bharathi.cse@sathyabama.ac.in

P.Jeyanthi, Information Technology, Sathyabama Institute of Science and Technology, Chennai, India. Email:jeyanthi.it@sathyabama.ac.in

M.Ramesh,TataConsultancyServices,Chennai,India,Email: ramesh.mohan@tcs.com

person patient. H. Zhang.et al[4] investigated the performance of Artificial Neural Network (ANN) models while applying to the survivability prediction on Chronic Kidney Disease (CKD) patients.

Dialysis or Kidney transplant stays the only option for a patient with End Stage Renal Disease (ESRD). The progression of the disease can be slowed down or even stopped in a favourable case by early prediction of CKD and proper treatments with diet. J. Aljaaf *et al*., [5] concluded that application of machine learning algorithms with predictive analytics proves to be an intelligent solution for early prediction of the disease.

Data mining models project ensemble techniques called Boosting which enhances the prediction of a model. AdaBoost and LogitBoost are generally used to compare the performance of classification algorithms. Arif-Ul-Islam. et al[6] analyzed the performance of boosting algorithms for detecting CKD and derived rules illustrating relationship among the various attributes of CKD. The paper used Ant-Miner machine learning algorithm along with Decision tree to derive rules.

Datamining methods are used to generate decisions by elicitating hidden information from chronic disease datasets. This calls of storage and manipulation of large amounts of structured, unstructured and semistructured data. The role of big data in for the same is very important. G. Kaur.et al[7] predicted chronic kidney disease using various data mining algorithms in Hadoop environment. Classifiers like KNN (K-Nearest Neighbor) and SVM (Support Vector Machine) are used in the research.

Levels of creatinine, sodium, urea in blood play an important role in deciding the survival prediction or the need for kidney transplantation in patients undergoing dialysis and becoming worser .V. Ravindra.et al[8] used simple K-means algorithm to elicit knowledge about the interaction between many of these CKD parameters and patient survival. He concluded that the clustering procedure predicts the survival period of the patients who undergo the dialysis procedure.

For CKD prediction R. Devika et al[9] examined the performance of Naive Bayes, K-Nearest Neighbour (KNN) and Random Forest classifiers based on accuracy, preciseness and execution time for. P.Panwong.et al[10] created a classification model for predicting transitional interval of Kidney disease stages 3 to 5 and also used Decision tree, K-nearest neighbor, Naïve Bayes and Artificial neural networks for eliciting the knowledge and creating classification model with the selected set of attributes.

S. Vijayarani.at.al[11] predicted kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN). The research compared the performance of the above two algorithms on accuracy and execution time. Misir R, et al[12] used feature selection algorithms to identify set of features that efficiently predict kidney diseases. The reduced feature set results in reduced costs, saves time and reduced uncertanity.

Kidney damage due to diabetes is chronic and a slow process, but has significant effects on the patient. High glucose levels in blood disturbs the kidney from functioning effectively. Bharathi et.al[16] has applied association rule mining to predict diabetes mellitus in a given dataset by generating summarisation rules. Revathy et.al[17,18] has used decision theory to validate the clustering data mining technique.

III. CKD PREDICTION USING MACHINE LEARNING MODELS

Knowledge discovery is an important application of datamining which involves various stages of processing. The application of datamining algorithms are facilitated by preprocessing the data collected from multiple sources. Data preperation or preprocessing involves cleaning, extracting and transforming data to suitable formats. The key factors of knowledge representation are identified from a larger feature set. Later various classification or pattern evaluation algorithms are applied for knowledge discovery. Bharathi [15] shows a general disease prediction model using machine learning.

The paper tries to propose a datamining framework for knowledge discovery on the CKD datasets. Large amounts of CKD datasets are collected. Data preperation and preprocessing is done using the traditional methods of data mining process. Three machine learning algorithms namely Decision tree, Random Forest and Support Vector machines are used to predict the early occurence of CKD. The goodness of each algorithm is analysed. The model with high accuracy is derived from the below process. The system architecture is given in Figure 1.

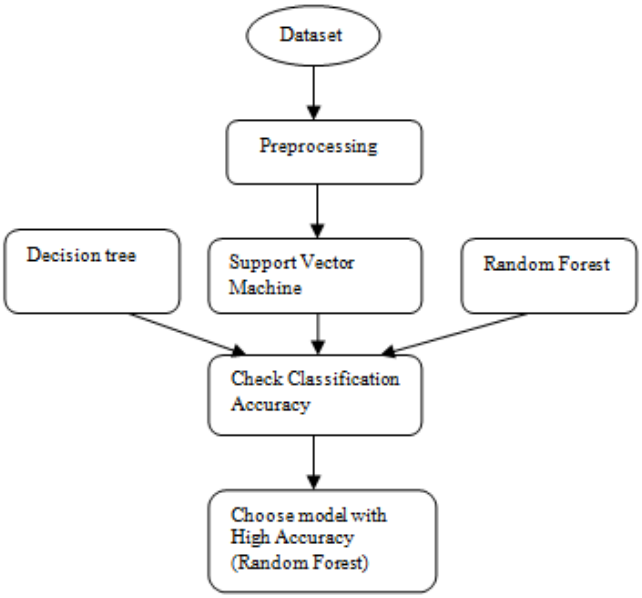

Fig.1.CKD prediction using machine Learning Models

A. Decision Tree

Decision Tree one of the algorithm, is used to solve regression and classification problems. The general objective of using Decision Tree is to create a model that predicts classes or values of target variables by generating decision rules derived from training data sets. Decision tree algorithm follows a tree structure with roots, branches and leaves. The attributes of decision making are the internal nodes and class labels are represented as leaf nodes. Decision Tree algorithm is easy to understand compared with other classification algorithms.

B. Support Vector Machine

A linear model for classification and regression is Support Vector Machine (SVM) that can be used to solve both linear and non linear problems. The algorithm classifies data using a hyperplane. In this algorithm, each data item will be plotted as a point in n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate. Classification will be performed by finding the right hyper-plane which can differentiate the two classes efficiently.

C. Random Forest

Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively fast and also accommodates missing data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees.

D. Classification Accuracy

Accuracy of the constructed classifier model can be calculation using the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where,

TP = Observation is positive an predicted is also positive
TN=Observation is negative and predicted is also negative
FP = Observation is negative but predicted is positive
FN = Observation is positive but predicted is negative

IV. DATA SET AND ALGORITHM

Experiments are conducted on Chronic Disease data set which has been downloaded from UCI machine Learning Repository [13] . This dataset contain 25 attributes(including target class attribute) and 400 instances which is shown in Figure 2. The most popular R Programming Data analytics tool has been used to construct the prediction framework.

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod
1	48	80	1.020	1	0	?	normal	notpresent	notpresent	121	36	1.2	?
2	7	50	1.020	4	0	?	normal	notpresent	notpresent	?	18	0.8	?
3	62	80	1.010	2	3	normal	normal	notpresent	notpresent	423	53	1.8	?
4	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111
5	51	80	1.010	2	0	normal	normal	notpresent	notpresent	106	26	1.4	?
6	60	90	1.015	3	0	?	?	notpresent	notpresent	74	25	1.1	142

Fig 2:Chronic Kidney Disease Dataset

Algorithm:

Input :Chronic Kidney Disease Dataset
Output: Hıgh Accuracy prediction Framework
Step1:Input data
Step2:Preprocess the data
Step 2.1:Convert Categorical values to numerical values
Step 2.2:Replace numerical missing values by Mean
Step2.3:Replace Categorical missing values by Mode
Step3:Construct Classifier Models
Step3.1:Construct Decision Tree Model
Step3.2:Construct Random Forest Model
Step 3.3:Construct SVM model
Step 4:Check the accuracy of the constructed models using confusion matrix.
Step 5:Decide the best prediction model for CKD.

V. RESULTS AND DISCUSSION

Models has been constructed using training data set(280 instances) which is 70% of original CKD data set. Constructed models have been validated using test data which is 30% of original data with respect to the parameter accuracy. Here , Accuracy has been calculated using confusion matrix .The best classifier model is the one with highest accuracy..

5.1 Accuracy of Decision tree

Confusion Matrix has been generated by decision tree model for the test data (120 instances)with class (values:CKD,NON CKD)as the target variable is given by table 1.The confusion matrix clearly says that 7 instances are not classified properly and 113 instances have been classified accurately and the accuracy of this classifier model is **94.16%.**

Table 1: Confusion Matrix –Decision Tree

	NON CKD	CKD
NON CKD	40	5
CKD	2	73

5.2 Accuracy of SVM

Confusion Matrix has been generated by SVM model for the test data(120 instances) with class (values:CKD,NON CKD)as the target variable is given by table 1.The confusion matrix clearly says that 2 insatcces are not classified properly and 118 instaces have been clasiified accurately and the accuracy of this classifiere model is **98.33%**

Table 2:Confusion Matrix –SVM

	NON CKD	CKD
NON CKD	42	2
CKD	0	76

5.3 Accuracy of Random Forest

Confusion Matrix has been generated by SVM model for the test data (120 instances) with class (values:CKD,NON CKD)as the target variable is given by table 1.The confusion matrix clearly says that 1 instances are not classified properly and 119 instances have been classified accurately and the accuracy of this classifier model is **99.16%.**

Table 3: Confusion Matrix –Random Forest

	NON CKD	CKD
NON CKD	42	1
CKD	0	77

5.4 Selection of Best Classifier Model

Accuracy of the various classifier models are listed in the table 4.It shows that Random Forest classifier model has provided highest accuracy.

Table 4: Accuracy of various classifier Models.

S.No	Classifier	Accuracy
1	Decision Tree	94.16
2	Support Vector Machine	98.33
3	Random Forest	**99.16**

VI. CONCLUSİON

This paper presented a prediction algorithm to predict CKD at an early stage. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters. Decision tree, Random Forest and Support Vector Machine learning models are constructed to carry out the diagnosis of CKD.The performance of the models are evaluated based on the accuracy of prediction. The results of the research showed that Random Forest Classifier model better predicts CKD in comparison to Decision trees and Support Vector machines. The comparision can also be done based on the time of execution, feature set selection as the improvisation of this research.

REFERENCES

1. M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," *2017 IEEE Life Sciences Conference (LSC)*, Sydney, NSW, 2017, pp. 300-303.
2. H. A. Wibawa, I. Malik and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2018, pp. 1-4
3. U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Chennai, 2016, pp. 1-5.
4. H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 1351-1356
5. J. Aljaaf *et al*., "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," *2018 IEEE Congress on Evolutionary Computation (CEC)*, Rio de Janeiro, 2018, pp. 1-9.
6. Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6.
7. G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," *2017 International* Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 973-979.
8. N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6.
9. V. Ravindra, N. Sriraam and M. Geetha, "Discovery of significant parameters in kidney dialysis data sets by K-means algorithm," *International Conference on Circuits, Communication, Control and Computing*, Bangalore,2014,pp.452-454.
10. R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 679-684.
11. P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," *2016 Second Asian Conference on Defence Technology (ACDT)*, Chiang Mai, 2016, pp. 145-150.
12. S. Vijayarani, S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS", International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.
13. Misir R, Mitra M, Samanta RK. A Reduced Set of Features for Chronic Kidney Disease Prediction. J Pathol Inform.2017.
14. "UCI Machine Learning Repository: Chronic_Kidney_Disease DataSet", Archive.ics.uci.edu, 2015. Available:http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
15. B.Bharathi, S.Prince Mary (2019), Neural Computation based general disease prediction model, International journal of Recent Technology and Engineering, vol8(2),pp 5646-5649.
16. Kamalesh M.D., Predictind the risk of diabetes mellitus to subpopulations using association rule mining, Proceedings of the International Conference of SoftComputing systems, Advances in Intelligent Systems and Computing col.397, Springer (2016)
17. Revathy,B.Parvathavarthini,Shiny Caroline, Decision Theory, an Unprecedented Validation Scheme for Rough-Fuzzy Clustering, International Journal on Artificial Intelligence Tools, World Scientific Publishing Company ,Vol.25,No.2,2016
18. Revathy Subramanion, Parvathavarthini Balasubramanian and Shajunisha Noordeen, Enforcement of Rough Fuzzy Clustering Based on Correlation Analysis,International Arab Journal of Information technology,IAJIT,Vol 14,No 1,91-98,2017.

AUTHORS PROFILE

Dr.S Revathy is a Associate Professor from the Department of Information Technology working in Sathyabama Institute of Science and Technology, Chennai India. Her research interest includes Machine Learning, Data Analytics and Big Data. She has published more than twenty papers in refereed journals.

Dr.B.Bharathi is a professor from the Department of Computer science and Engineering working in Sathyabama University, Chennai, India. Her major areas of research include software engineering, software architectures, machine learning and Big Data. She has more than 50 publications in refreed journals. She is the reviewer of many refereed journals and also acted as advisory member for various conferences.

P Jeyanthi is a Associate Professor from the Department of Information Technology working in Sathyabama Institute of Science and Technology, Chennai India. Her research interest includes Image Processing and Machine Learning. She has published more than ten papers in refereed journals.

M Ramesh has received M.C.A from Bharathidasan University. Currently, he is working as a Associate Consultant in Tata Consultancy Services. His research interest includes Devops, Machine Learning and Data Analytics.