Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms

Shanila Yunus Yashfi
Dept. of Computer Science and Engineering
North Western University.
Khulna, Bangladesh
shanilayunus54@gmail.com

Md Ashikul Islam
Dept. of Computer Science and Engineering
North Western University.
Khulna, Bangladesh
ashik.nwu@gmail.com

Pritilata
Dept. of Computer Science and Engineering
North Western University.
Khulna, Bangladesh
pritilatabd@gmail.com

Nazmus Sakib
Dept. of Computer Science and Engineering
North Western University.
Khulna, Bangladesh
nazmussakibnwu@gmail.com

Tanzila Islam
Dept. of Computer Science and Engineering
North Western University.
Khulna, Bangladesh
tanzeelaislam607@gmail.com

Mohammad Shahbaaz
Dept. of Electrical and Computer Engineering
Memorial University of Newfoundland.
St.John's, Canada
Mshahbaaz@mun.ca

Sadaf Salman Pantho
Dept. of Computer Science and Engineering
Jashore University of Science and Technology.
Jashore, Bangladesh
sadafpantho@gmail.com

***Abstract*—CKD is a serious reason of demise and disability. It was the 27th focal reason in 1990 and became 18th focal reason in 2010. Near about 1 million people lose their life in 2013. In spite of that, people of developing countries are being affected by CKD. We analyzed the data of CKD patient and proposed a system from which it will be possible to predict the risk of CKD. We have used 455 patients' data. Online data set which is collected from UCI Machine Learning Repository and real time dataset which is collected from Khulna City Medical College are used here. We used Python as a high-level interpreted programming language for developing our system. We trained the data using 10-fold CV and applied Random forest and ANN. The accuracy achieved by Random forest algorithm is 97.12% and ANN is 94.5%. This system will help to predict early disclosure of chronic kidney diseases.**

Keywords—Chronic Kidney Disease, Random Forest Algorithm, Artificial Neural Network.

I. INTRODUCTION

The kidneys are a two of a kind of organs placed towards the lower back of the abdomen. Kidneys job is to strain the blood by moving out the toxic substance from the body using the bladder through urination. Kidney failure can cause death if the kidneys do not remove waste which is affected toxins. Difficulties of Kidney can be classified as acute or chronic. Chronic kidney diseases include circumstances that harm kidneys and reduce their capability to keep us fit. If kidney problem gets worse, the waste can build up at zenith levels in our blood and can cause difficulties such as hypertension, weak bones, anemia, bad nutrition and nerve injure. Furthermore, kidney disease raises the possibility of heart and vascular diseases.

CKD can be caused by diabetes, hypertension, coronary heart disease, lupus, anemia, bacteria and albumin in the urine, complications of some drugs, sodium and potassium deficiency in the blood and a family history and many others. Early disclosure and medical care can usually avoid the worsening of chronic kidney disease. If it can get worse it can result in to kidney failure requiring dialysis. It can also result for kidney transplantation to sustain living.

Chronic kidney disease (CKD) is a universal public wellbeing difficulty. 10% of the world's population is affected by CKD. However, there is little direct evidence on how to systematically and automatically diagnose CKD. Worldwide, CKD is a serious reason of demise and disability. It was the 27th focal reason in 1990 and became 18th focal reason in 2010 [1]. Near about 1 million people lose their life in 2013 [2]. In spite of that, people of developing countries are being affected by CKD. In 2015 an orderly planned analysis, conducted and proclaimed that in high-income countries 109.9 million people had CKD in which 48.3 million and 61.7 million people were men and women respectively while the load was 387.5 million in countries with low average incomes where men were 177.4 million and women were 210.1 million [3]. According to the Kidney Foundation, out of 18 million people about 35,000 to 40,000 patients suffer from with chronic renal failure in Bangladesh each year.

Director of the National Institute of Renal Diseases and Urology (NIKDU) Nurul Huda Lenin said BSS that, women are more comparing to men in case of by CKD. Referring to the data on "Global prevalence of CKD: a systematic review and meta-analysis", he stated that the 14% of people are women while 12% people are men and still men percentage are high during dialysis. According to Professor Lenin there are three main causes which are responsible for CKD among women: neglect, social barriers and lack of consciousness. Professor Harun Ur Rashid, president of Bangladesh Kidney Foundation said that the patients of diabetes and hypertension are increasing in Bangladesh for different reasons, including poor food habit and undisciplined daily life which lead them kidney disease.

The motive of our research is to scrutinize the CKD data and portend the risk of CKD using Random Forest Algorithm and Artificial Neural Network.

II. Background Study

S.Gopika, et al. [4] has performed a method of CKD Prediction with Clustering Method. The main objective is to determine the kidney function failure by using clustering algorithm. The experimental outcome revealed that the Fuzzy C means algorithm renders superior results and its accuracy 89%. Similarly Tabassum S, et al. [5] has explored big data in healthcare which has been developed by conducting data mining techniques. EM is used to cluster parallel type of individual into one set. ANN and C4.5 are classification method which is used for prediction of the disease. EM got the accuracy result of 70% which is a type of clustering algorithm. ANN and C4.5 is classification algorithm in which ANN got the accuracy result of 75% and C4.5 algorithm got the accuracy result of 96.75%. Sujata Drall, et al. [6] performed an innovative method by conducting Data Mining, Machine learning and different classification algorithms which is focused on predicting CKD status of a sick person with high accuracy. In this research they have used 5 CKD attributes out of 25 and two classification algorithms KNN and Naïve Bayes for portending CKD of a patient. KNN classifier predicted CKD with an accuracy of 100%, whereas Naïve Bayes Classifier predicted with an accuracy of 96.25%. Siddheshwar Tekale, et al. [7] has analyzed 14 dissimilar characteristics correlated to CKD sick person and anticipated accuracy for several ML procedures alike Decision tree and Support Vector Machine. It is detected from the results study that decision tree algorithms gives the accuracy of 91.75% and SVM renders accuracy of 96.75%. The prediction process is less time consuming. Asif Salekin, et al. [8] has considered 24 analytical factors and generates a ML classifier to predict CKD. They evaluated their method on a dataset of 400 entities, wherever 250 were CKD and they achieved a prediction accuracy of 0.993. They additionally performed feature assortment to choose the utmost appropriate features for recognizing CKD and flourishing them consistent with their certainty. Faisal Aqlan, et al. [9] has showed a DST that assist in the analysis of CKD. Data mining and analytics methods can be designed for detecting CKD by conducting chronological patient's report and analysis proceedings. In this research DT, LR, NB and ANN were intended for detecting CKD. Random trees give 100% accuracy. El-Houssainy A. Rady, et al. [10] have experienced that Effective data mining strategies is appeared to uncover besides concentrate concealed data The outcomes of relating PNN, MLP, SVM and RBF procedures have been associated. PNN process gives the maximum accuracy of 99.7%, as related with all other algorithm consequences. Sahil Sharma, et al. [11] has applied various machine learning algorithms for CKD. 400 cases and 24 features are used for the research. The outcomes demonstrate that DTdidthe best results with the accuracy of 98.6%, sensitivity of 0.9720, precision of 1 and specificity of 1. Pratibha Devishri S, et al. [12] had demonstrated here Feature selection approach is suitable for CKD prediction. They examined Decision stump, Rep tree, IBK, K-star, SGD and SMO classifier using WEKA. Accuracy measures used to compare classifiers are Recall, F-measure and Precision by implementing on WEKA. They got 87.3% accuracy. GUNEET KAUR, et al. [13] has introduced the CKD prediction with data mining algorithms. Here KNN and SVM data mining procedures are accustomed forecast the CKD. Six parameters are produced from the dataset which are Accuracy, Error, Precision, Recall, F1 error, Elapsed Time. MATLAB tool used to perform for prediction of CKD by accessing Hadoop in itself.

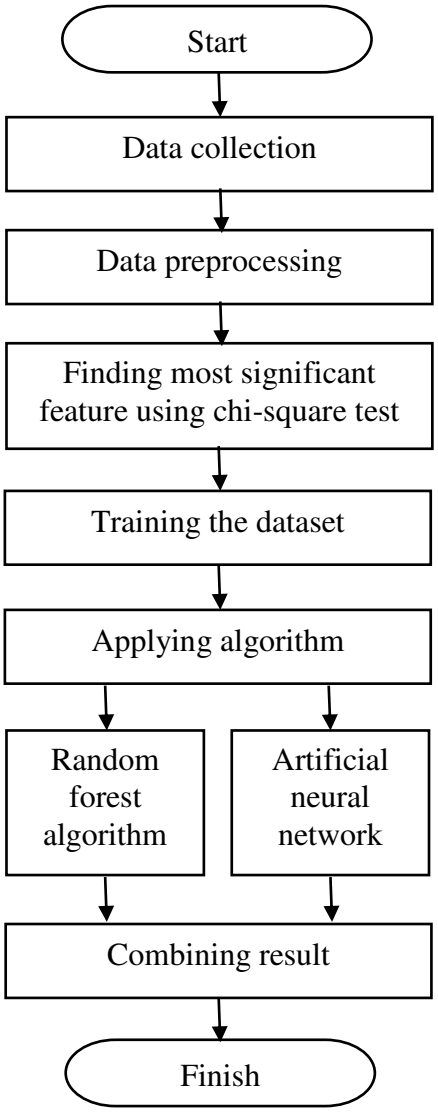

Fig. 1. System structure of proposed method

In this study, we worked with 55 real life dataset and used chi-square test which is a feature selection algorithm. As mentioned earlier, we used ANN and Random forest algorithm. The uniqueness of our study is, we merged the result of the two algorithms using logic mentioned below and predicted the risk.

III. Research Methodology

This section illustrates the entire concept of the research work which will aid to understand the whole notion of the paper. At first, we collected the data and preprocess it. After preprocessing the data, we handled the missing data of the dataset using Anaconda. Feature selection is conducted to extract the most significant features. Then, we apply ANN and Random Forest algorithm on the dataset. Figure 1 shows the system structure of proposed method. We have divided our thesis work in the following orders:

A. Data Collection

We collected data in two ways. Online data set which is collected from UCI. The amount of data collected from UCI is 400 instances each with 25 features Real time data set is collected from Khulna City Medical College and the amount of data is 55 instances each with 25 features. Features are same in Online data set and Real time data set.

B. Data

It is an essential step to gain better accuracy. Before going out in a journey it is really important to be prepared. It is same for the machine learning journey. If there will be no data pre-processing than machine learning model won't work appropriately. After collecting online data set and real time dataset we will proceed for data processing. “Pandas” and “Numpy” library is used for processing the data.

C. Handling Missing Data

While processing data, there were so many incomplete data means there were numerous missing data which occurs myriad of time in real life and we managed the missing data using “Median” method.

D. Extracting The Most Significant Feature Using Chi-Square Test

There were few features in the dataset which was not relevant to the objective. So feature selection is a vital step. By using chi-square test we extracted the most significant features. There were 24 features excluding class feature. After extracting there are 20 features and then we trained the dataset.

E. Applying 10-Fold Cross Validation

10-Fold cross validation is a technique where dataset is divided into 10 numbers of folds and each and every data set is used as a testing data set at some point and it will not halt until 10 folds have been used as a testing set.

F. Applying Algorithms

We used two classification algorithms. The algorithms are: Random Forest and ANN. Random Forest Algorithm is an ensemble classifier which used decision tree algorithm in a randomized way. The more decision trees the more robust prediction and the higher accuracy. For decision tree, we have a dataset from this dataset we selected the target attribute. Then find the Information Gain of the target attribute. For remaining attributes we estimated the value of Entropy and then found out their Information Gain by multiplying their Entropy with their probability. Last step is to find out each attribute Gain which is the difference between Information Gain (IG) and Entropy E (A). The attribute with the highest gain will be the decision tree root node. Every decision tree predicted a result. We used random forest classification for our proposed work. The value with the maximum numbers of vote is the predicted result of random forest algorithm.

Artificial neural network (ANN) is a supervised learning method. It is made of a huge number of simple elements, called perceptrons. A neural system has just three layers of neurons, an input layer that receives the input, one hidden layer and an output layer that produces output as shown in figure 2. A multilayer perceptron (MLP) is a gathering of perceptrons, sorted out in various layers that precisely respond to composite questions. Each and every perceptron emits signals from input layer to hidden layer. An MLP constitute one layer which is named as input layer another one which is named as hidden layer, and at last output layer which is shown in Figure 2 [14].

Here we used one input layer, three hidden layers and one output layer. 800 iterations happened here with 24 independent variables. Back propagation algorithm is a group of techniques which is used to instruct ANN. ANN is following a gradient-based optimization algorithm that

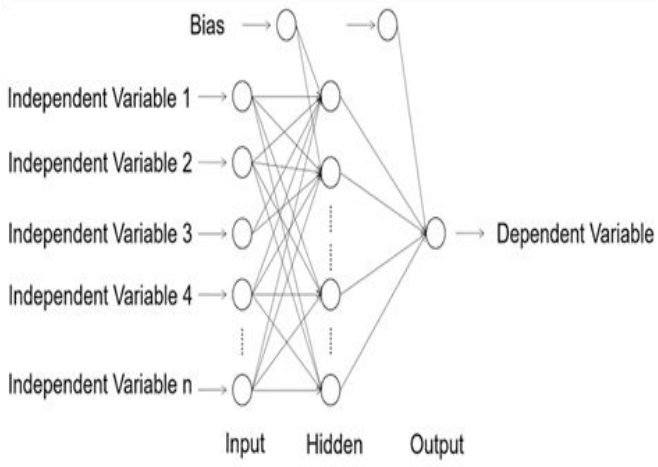

Fig. 2. The Perceptron Learning Process

follows the chain rule. The primary component of back propagation is its recurrent, recursive and effective technique for computing the weight updates to improve the system until [15].

G. Combining Results

There were few features in the dataset which was not relevant to the objective. So feature selection is a vital step. By using chi-square test we extracted the most significant features. There were 24 features excluding class feature. After extracting there are 20 features and then we trained the dataset.

IV. Experimental Results

TABLE I. Feature Name and Feature Type of the Chronic Diseases Dataset

SL	Feature	Data type
1	Age	Numerical
2	Blood Pressure	Numerical
3	Specific Gravity	Numerical
4	Albumin	Numerical
5	Sugar	Numerical
6	Red Blood Cell	Nominal
7	Pus Cell	Nominal
8	Pus Cell Clumps	Nominal
9	Bacteria	Nominal
10	Blood Glucose Random	Numerical
11	Blood Urea	Numerical
12	Serum Creatinine	Numerical
13	Sodium	Numerical

14	Potassium	Numerical
15	Hemoglobin	Numerical
16	Packed Cell Volume	Numerical
17	White Blood Cell Count	Numerical
18	Red Blood Cell Count	Numerical
19	Hypertension	Nominal
20	Diabetes Mellitus	Nominal
21	Coronary Artery Diseases	Nominal
22	Appetite	Nominal
23	Pedal Edema	Nominal
24	Anemia	Nominal

TABLE II. OUTCOMES OF RANDOM FOREST ALGORITHM

Evaluation Metrics	Random Forest Algorithm
Total number of instances	455
Accuracy	97.12%
Error	2.88%
Recall (Weighted Avg.)	0.97
F1-score (Weighted Avg.)	0.97
Precision (Weighted Avg.)	0.97

The accuracy has received for random Forest is 97.12%. Error occurred 2.88%. Precision, Recall and F-Measure are 0.97.

TABLE III. OUTCOMES OF ARTIFICIAL NEURAL NETWORK

Evaluation Metrics	Artificial Neural Network
Total number of instances	455
Accuracy	94.5%
Error	5.5%
Recall (Weighted Avg.)	0.94
F1-score (Weighted Avg.)	0.95
Precision (Weighted Avg.)	0.95

The accuracy has received for artificial neural network is 94.50%. Error Occurred 5.5%. Precision, Recall and F1 score are 0.95, 0.94 and 0.95 respectively.

TABLE IV. COMBINING RESULTS

Random Forest	ANN	Output
0	0	LOW
0	1	MEDIUM
1	0	MEDIUM
1	1	HIGH

V. CONCLUSION AND FUTURE PROSPECT

To prevent any type of disease the most significant step is to detect the disease which is expensive in country like Bangladesh. Because of this, people in Bangladesh suffer the most. Nowadays, CKD patient rate in Bangladesh are increasing rapidly. So our aim is to forge a system which will help people to predict the risk of chronic kidney disease. In our proposed work, we have used UCI dataset and real time dataset and processed them. We have handled missing data, trained it and made a Random Forest and ANN model. We have implemented these two algorithms in python language. The accuracy, we gain using Random Forest algorithm is 97.12% and ANN is 94.5% respectively which is relatively very good. By using our proposed method, risk prediction of CKD in early stage will be possible. Collecting real time dataset is a difficult work. At the time of collecting real time data, we have faced several problems. After collecting data there were another problem is to handle missing value which was handled. The reason behind development and success is idea. Even we have some idea which will make our proposed work more efficient. Nowadays people from low to high income uses a mobile phone. So we intend to build a mobile application with our proposed work. Because of that, people will be able to use the application which will be easy for them to detect CKD. We have used only one filter method which is Chi-square test in our work. In future, we will also use wrapper method to extract significant features more accurately. Most of our data are from online. In future we intend to work with real time dataset and achieve higher accuracy.

REFERENCES

[1] "Global Facts: About Kidney Disease," National Kidney Foundation, 2015.[Online].Available:https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease. [Accessed: 20-Sep-2019].

[2] "Over 35,000 Develop Kidney Failure In Bangladesh Every Year." Thedailystar.net,2019.[Online].Available:https://www.thedailystar.net/city/news/18m-kidney-patients-bangladesh-every-year-1703665. [Accessed: 20-Sep-2019].

[3] "Women More Affected By Kidney Diseases", en.prothomalo.com, 2018.[Online].Available:://en.prothomalo.com/bangladesh/Women-more-affected-by-kidney-diseases. [Accessed: 20-Sep-2019].

[4] S. Gopika, and Dr. M. Vanitha, "Machine learning Approach of Chronic Kidney Disease Prediction using Clustering Technique" ,*International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 6, no. 7, pp.14488-14496, 2017.

[5] Tabassum, Mamatha Bai B G, Jharna Majumdar, "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques", *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE),* vol. 4, no. 9, pp.25-31, 2017.

[6] Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib, "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach", *International Journal of Management, Technology And Engineering*, vol. 8, no. 5, pp.278-287, 2018.

[7] SiddheshwarTekale, PranjalShingavi, Sukanya Wandhekar, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm",*International Journal of Advanced Research in Computer and Communication Engineering,* vol. 7, no. 10, pp.92-96, 2018.

[8] Asif Salekin, Jhon Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes", in 2016 IEEE International Conference on Healthcare Informatics (ICHI), Chicago, USA, 2016, pp. 262-270.

[9] Faisal Aqlan, Ryan Markle, Abdulrahman Shamsan, "Data mining for chronic kidney disease prediction", 67th Annual Conference and Expo of the Institute of Industrial Engineers, United States, 2017, pp. 11789-1794.

[10] El-Houssainy A. Rady, Ayman S. Anwar, "*Prediction of Kidney Disease Stages using Data Mining Algorithms*", Informatics in Medicine Unlocked (IMU 100178), Article Number: 100178, 2019.

[11] Sahil Sharma, Vinod Sharma, and Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis", *International Journal of Modern Computer Science (IJMCS)*, vol. 4, no. 3, pp.11-15, 2016.

[12] Pratibha Devishri S, Ragin O R, Anisha G S, “Comparative Study of Classification Algorithms in Chronic Kidney Disease”, *International Journal of Recent Technology and Engineering (IJRTE),* vol.8, no.1, pp.180-184, 2019.

[13] Guneet Kaur, Er.Ajay Sharma , “Predict chronic kidney disease using data mining algorithms in hadoop “, in 2017 Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), 2017, pp. 973-979.

[14] Hao Li,Zhien Zhang, ZhijianLiu,"Application of Artificial Neural Networks for Catalysis: A Review", *Multidisciplinary Digital Publishing Institute,* vol. 7, no. 10, pp. 2-19, 2017.

[15] "Complete Guide to Artificial Neural Network Concepts & Models", MissingLink.ai.(2019).[Online].Available:https://missinglink.ai/guides/neural-network-concepts/complete-guide-artificial-neural-networks. [Accessed: 30-Sep-2019].