Chronic Kidney Disease Prediction using Neural Network and ML Models

Meghana H L
Department of Computer Science
B.M.S College of Engineering
Bangalore, India

Vaishnavi S Kuber
Department of Computer Science
B.M.S College of Engineering
Bangalore, India

Yamuna B S
Department of Computer Science
B.M.S College of Engineering
Bangalore, India

Varshitha T L
Department of Computer Science
B.M.S College of Engineering
Bangalore, India

Prof. Vikrant. B. M
Department of Computer Science
B.M.S college of Engineering
Banglore, India

***Abstract*—In Today's world, everyone is conscious of health. Because of the most dominant IT lifestyle, the workload is more and people hardly give attention to health unless it turns worse. Chronic kidney disease is a kind of disease that hardly shows symptoms in the early stages and in later stages, things become worse which might end in kidney failure or an artificial support system. Thus our system aims at predicting the disease early and will help in taking precautionary measures or early medication. We use three algorithms and analyze the performance of them. The algorithms used are support vector machine, random forest, and a hybrid neural network model**

Keywords*—Chronic Kidney Disease; hybrid model; Random Forest; Support Vector Machine;

I. INTRODUCTION

The kidney is a very important organ in the human body whose main functions include osmoregulation and excretion. All the harmful and unwanted materials from the body are gathered and excreted by the kidney and excretory system. In India, every year there are around 1 million cases of Chronic Kidney Disease (CKD), also called renal failure. It can cause a loss in kidney functionality and hence it can pose lots of danger. CKD is a rather slow and timely loss of kidney function over a very long duration of time. A person can develop permanent kidney failure. If one fails to diagnose CKD at an earlier stage then the patient can show symptoms of weak bones, anemia, nerve damage, etc. Therefore it is the need of the hour to detect this disease in the early stages but since the symptoms become unpredictable and at times are not specific to the disease. Thus for diagnosis in people who show no symptoms at all machine learning can be of help. It uses an old CKD patient data set to train models.

The papers reviewed below show the various techniques and algorithms used in machine learning to predict chronic kidney disease. We aim to predict the disease using a hybrid model and make the prediction more efficient.

II. LITERATURE SURVEY

Chronic kidney disease has become one of the main concerns in health care. Chronic kidney disease is a lasting disease it keeps getting worse over time. It may also lead to kidney failure. Some of the symptoms like nausea, vomiting, muscle cramps, appetite losses, swelling in the feet, ankles indicate kidney disease[5]. Machine learning can be used to find out the presence of the disease. Various classification models and data mining techniques have been used in the past. In any machine learning problem pre-processing of the data before applying any algorithm. Data mining statistical techniques are used for data pre-processing. Mode, median methods are used for nominal values whereas mean is for numerical values[5,13,3,10]. Classification and detection accuracy of pre-processing using mean mode, median with neural networks give better results than SVM, KNN(K-Nearest Neighbor), regression tree[5]. Sometimes tuples having missing values are excluded which can be done using the WEKA(Waikato Environment for Knowledge Analysis) function[1,2,6].

In[2] research paper the number of features for the prediction was reduced using the correlation-based feature selection method. Then the dataset was used to train incremental backpropagation learning networks and Levenberg–Marquardt. A reduced number of attributes helps in reducing uncertainty in decision-making.It enhances the comprehensibility of data, facilitates better visualization of data, reduces training time of learning algorithms, and improves the performance of prediction[4]. It is a part of pre-processing[4]. There are three traditional feature selection categories. They are filter, wrapper, and embedded method[4]. Filter methods are best for large datasets over wrapper methods. Filter methods are applied before applying any classifier hence they are independent of the classifier. This makes them fail to select useful features[4]. Wrapper methods are better than filter methods in selecting the most useful features but they need a large dataset otherwise there is a problem of overfitting. Embedded methods are usually guided by the learning process. They usually work according to a specific learning algorithm which helps in optimizing the performance of a learning algorithm. This method makes better usage of available data and provides faster solutions as they do not require splitting of training data into the training set and validation set. They are computationally inexpensive and less prone to overfitting compared to wrapper methods. Hybrid methods are used to take the advantage of different methods to optimize results[4].

Both data mining and classification techniques are applied in chronic kidney disease prediction. Models like SVM, Decision tree, K-NN, Naive Bayes, neural networks are used for the prediction of the diseases[1,14,]. In[3] random forest gave the best accuracy(99%) compared to decision tree and SVM. In[7] Eleven techniques of decision trees like decision Stump, J48, CTC, LMT, NBTree, Random Forest, randomTree, REPTree, simple Cart, J48graft are applied to the dataset. Random forest outperforms other methods. Neural networks with three layers performed better compared with other models like SVM, DT, KNN, Gradient Boost[1]. 10-fold cross-validation technique was used which gave the neural network 98.25 accuracy[1]. In[5] a multilayer neural network with back-propagation is used for the prediction. It is also mentioned that the complexity of the neural network increases if the number of hidden layers increases, which may increase accuracy and computational time. Artificial neural network outperforms SVM(Support Vector Machine), classification tree, KNN, Regression tree.

Five missing imputation techniques like fixed using mean, fixed using mid-range, random uniform, random normal, and Classification and Regression Trees (C&RT) algorithm are used if a variable has less than 15% missing data otherwise it is removed. k-means clustering and regression can be used to predict the stages of the disease. The randomTree gave the best results compared to Neural Networks, Logistic Regression, Bayes Net, Chi-square Automatic Interaction Detector (CHAID), and Support Vector Machines[10].

When the Gravitational search algorithm is used with an artificial neural network, it enhances the classification accuracy of the neural network by skipping local minima and converging to global minima. Genetic algorithms with artificial neural network gave more accuracy compared to the neural network and KNN[12].

A deep neural network with three layers is used in[11]. Relu and tanh activation functions gave the best results. The sigmoidal activation function is used in the output layer. The deep neural network outperformed compared with other models like SVM, Naive Bayes, adaboost, logistic regression, random forest[11].

In[13] Density-Based feature selection with Ant Colony based Optimization technique is used for kidney disease prediction. Density based feature selection is a heuristic method for finding feature importance. Proposed model outperformed earlier models.

In[14] the Bayes theorem is implemented using Naïve Bayes Classifier with assumptions which would deal with conditional basic probabilities. Decision trees employed in data mining are classification tree analysis and Regression tree analysis. For varied predicted output such as belonging to particular data class or real number. Usage of Data mining technique for different analysis of medical data is a better method. The performance of Decision tree method is 91% accurate when compared with naive Bayes method. Data Mining helps to obtain correlations from attributes which are not direct indicators of the class.

In[9] hybrid neural network is used for analyzing both text and properties on EHR(Electronic Health Records) collected from across the country in China. Bidirectional Long Short_Term memory(BiLSTM) for text analysis and autoencoder network for properties are used. The hybrid neural network is more efficient than statistical methods[9]. A polynomial support vector machine is used in[8] which gives the most accurate results.

Various performance parameters like accuracy, specificity, precision, sensitivity, false-positive rate, F-measure are used in health care to analyze the performance of the model[4,8].

III. PROPOSED APPROACH

In our approach, we pre-process the data and select the most important features. Only those features are selected and the rest are eliminated. The model is trained and tested using various models like SVM, Random forest and hybrid neural network model. The efficiency of various algorithms are compared using different performance parameters.

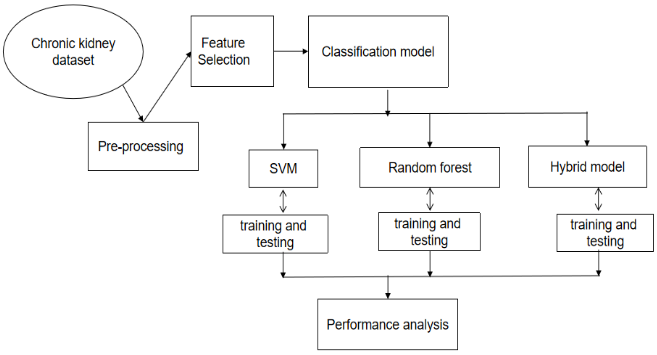

Fig. 1. System Architecture

Our approach has four steps: pre-processing, training the model, testing the model, evaluating performances.

1. *A. Pre-process*

The dataset is taken from the Kaggle website, all the data is collected from the hospital in India as per it is mentioned on the website. Dataset has features like blood sugar, albumin, red blood cells, puss cells, blood pressure etc. Dataset has 400 rows and 26 columns. Id column which uniquely identifies each entry in the dataset is removed. All the records or rows containing null values are removed. Categorical values in the dataset are all replaced by binary number 0 or 1.

Visualisation of the dataset shows that aged people and people with blood pressure are more prone to CKD.

2. *B. Feature Selection*

Extra tree classifier is used to select the most important features. It is very similar to the random forest but it selects the splits randomly whereas random forest selects optimal split. In the extra tree classifier, multiple de-correlated decision trees are constructed. They are constructed by selecting a random set of features. While constructing forest normalized reduction of mathematical features is calculated for each of the features. If Gini index is used, it is computed for each feature and it acts as feature importance of the respective feature. The greater the gini index, the more important the feature is. Feature importance of each of the features is found out based on the selected threshold 18 features are selected.

3. *C. SMOT(Synthetic Minority Oversampling technique)*

SMOT technique is used to oversample minority class in the dataset in classification problem. In our dataset, we have more patient records having CKD disease than non-CKD. Classifier's performance reduces because of imbalanced classes. Hence minority class is duplicated by the SMOT technique before giving data to the classifier.

4. *D. Dividing the data*

20 percent of the data is used for testing and 80 percent for training the model.

5. *E. Random Forest*

After hyperparameter tuning the best parameters for constructing a random forest is found out. The model is fitted with the training data. Both testing and training accuracy are 100% which shows that the model is overfitting as the training data is less.

TABLE I. CLASSIFICATION REPORT OF RANDOM FOREST

	precision	recall	f1-score	support
0	1.00	1.00	1.00	27
1	1.00	1.00	1.00	19
accuracy			1.00	46
macro avg	1.00	1.00	1.00	46
Weighed avg	1.00	1.00	1.00	46

6.

7. *F. SVM(Support Vector Machine)*

After hyper parameter tuning best parameters for SVM are found out. SVM is fitted with training data. Testing accuracy is 82.60%, training accuracy is 98.36%.

TABLE II. CLASSIFICATION REPORT OF SVM

	precision	recall	f1-score	support
0	0.92	0.81	0.86	27
1	0.77	0.89	0.83	19
accuracy			0.85	46
macro avg	0.84	0.85	0.85	46
Weighed avg	0.86	1.00	0.85	46

8. *G. Neural network*

A cluster of layers are used which are linearly stacked to create a model. Convolution layer is added as a first layer with ReLU activation function.

ReLU function (1) returns the output directly if the given input is positive returns zero otherwise.

$$Y = \max(0,X) \qquad (1)$$

Y=output of the ReLU function
X=input to the ReLU function

Second layer is max pooling layer, followed by Long Short Term Memory , which is followed by a dense layer which is regular deeply connected neural network layer with ReLU activation function. Last layer is dense with softmax function.

Softmax function (2) takes input values as vector and returns probability values which sum up to one.

$$\sigma(Z)_j = \frac{e^{Z_i}}{\sum_{j=1}^{k} e^{Z_j}} \qquad (2)$$

σ = softmax
Z= input vector
e^{Z_i} = standard exponential function for input vector
e^{Z_j} = standard exponential vector for output vector
k = number of classes

Accuracy of hybrid neural network is 97.82% .

TABLE III. CLASSIFICATION REPORT OF NEURAL NETWORK

	precision	recall	f1-score	support
0	0.90	1.00	0.95	18
1	1.00	0.93	0.96	28
accuracy			0.96	46
macro avg	0.95	0.96	0.96	46
Weighed avg	0.96	0.96	0.96	46

IV. CONCLUSION

We selected the most important features using Extra Tree Classifier. When Random forest was applied for the dataset it gave 100% accuracy which indicates that the model is overfitted. The accuracy of SVM is 84.78% and of hybrid model is 97.82%. The hybrid model outperformed both SVM and Random forest. Even with a small amount of data, a hybrid neural network shows better performance compared to SVM and Random forest. Performance of the ML models can be improved in future by training with large amounts of data.

ACKNOWLEDGMENT

We would like to thank Professor Vikrant.B.M for guidance throughout our study. We also thank B.M.S college of engineering for supporting us to complete this study.

REFERENCES

[1] Bandyopadhyay, Samir Kumar, and Shawni Dutta. "Chronic kidney disease prediction using neural approach." medRxiv (2020).

[2] Misir, Rajesh, Malay Mitra, and Ranjit Kumar Samanta. "A reduced set of features for chronic kidney disease prediction." *Journal of pathology informatics* 8 (2017).

[3] S Revathy, B Bharathi, P. Jeyanthi, M Ramesh "Chronic kidney disease prediction using machine learning models." International journal of engineering and advanced technology (volume 9 issue-1 october 2019).

[4] Jain, Divya, and Vijendra Singh. "Feature selection and classification systems for chronic disease prediction: A review." Egyptian Informatics Journal 19, no. 3 (2018): 179-189.

[5] Chakrapani, Sumitraj, Vibhavprakasha Singh, Dhrubjyoti Kalita " Detection of chronic kidney disease using artificial neural networks." International Journal of applied engineering research(Volume 14, November 10 2019).

[6] Pranjal shingavi, Sukanya wandekar, Ankit Chatotikar " Prediction of chronic kidney disease using machine learning algorithm." International Journal of advanced research in computer and communication engineering(Vol 7, Issue 10, October 2018).

[7] Pasadana, I. A., D. Hartama, M. Zarlis, A. S. Sianipar, A. Munandar, S. Baeha, and A. R. M. Alam. "Chronic kidney disease prediction by using different decision tree techniques." In Journal of Physics: Conference Series, vol. 1255, no. 1, p. 012024. IOP Publishing, 2019.

[8] Mohamed Alloghani, Abir Hussain, Dhiya AI -Jumeliy, "Performance -based prediction of chronic kidney disease using machine learning for high-risk cardiovascular disease patients.", Springer nature Switzerland,September-2019.

[9] Ren, Yafeng, Hao Fei, Xiahui Liang, Donghong Ji, and Ming Cheng. "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records." BMC medical informatics and decision making 19, no. 2 (2019): 51.

[10] Aqlan, Faisal, Ryan Markle, and Abdulrahman Shamsan. "Data mining for chronic kidney disease prediction." In IIE Annual Conference. Proceedings, pp. 1789-1794. Institute of Industrial and Systems Engineers (IISE), 2017.

[11] Kriplani, Himanshu, Bhumi Patel, and Sudipta Roy. "Prediction of chronic kidney diseases using deep artificial neural network technique." In Computer Aided Intervention and Diagnostics in Clinical and Medical Images, pp. 179-187. Springer, Cham, 2019.

[12] Pooja Sharma, Swarnadeep Saket, "Survey for the Prediction of Chronic Kidney Disease using Machine Learning" at al Int J Sci Res Sci Eng Technol, December-2019.

[13] Elhoseny, Mohamed, K. Shankar, and J. Uthayakumar. "Intelligent diagnostic prediction and classification system for chronic kidney disease." *Scientific reports* 9, no. 1 (2019): 1-14.

[14] Padmanaban, K. Anantha, and G. Parthiban. "Applying machine learning techniques for predicting the risk of chronic kidney disease." Indian Journal of Science and Technology 9, no. 29 (2016): 1-6.