CHRONIC KIDNEY DISEASE PREDICTION USING CGAN

Monika Vishnuvarthini A[1] , Dr. C. Thirumarai Selvi[2] , Sowmiya. M[3] and Shyam Kumar[4].
[134]Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India
[2]Professor, Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India

Abstract. Chronic kidney disease (CKD), is also known as chronic renal disease. Chronic kidney disease means lasting damage to kidneys if the disease in a very bad condition then our kidney may stop working and also decrease our health. By, this we may have complications like high blood pressure, anemia (low blood count), weak bones, poor nutritional health, and nerve damage. To avoid these conditions we should do early detection and treatment for Chronic Kidney Disease. The objective of this research work is to introduce a new deep learning decision support system to predict chronic kidney disease. This work aim is to give a better performance through Conditional Generative Adversarial Network (CGAN) classifier based on its accuracy, precision, and recall for CKD prediction. From the experimental results, it is observed that CGAN gives better performance while comparing with K-nearest neighbor (KNN) classifier.

Keywords: Chronic kidney disease, K-nearest neighbour, missing value elimination, recursive feature elimination, Conditional generative adversarial network.

1 INTRODUCTION

Chronic kidney disease (CKD) is a collection of heterogeneous disorders in kidney which is associated with gradual loss of kidney function. The term CKD depicts a lasting damage in the kidney that it can be also termed as Kidney failure or end-stage renal disease (ESRD). Kidney being an important and critical excretory organ if failed an external aid like hemodialysis or even a kidney transplant to survive. People can get CKD irrespective of their age while some are more at risk than others like people with Diabetes, hypertension, heart disease, etc., are more prone to the disease. There are some symptoms through which kidney failure can be predicted which includes itching, muscle cramps, not feeling hungry, swelling in your feet and ankles, and an excessive amount of urine or not enough urine, etc.,

Studies suggest that early detection of CKD can potentially prevent worsening of kidney that leads to dialysis or transplantation. If a person is suspected to have CKD then either a kidney imaging or a datasets with bp, sugar, RBC etc are used to detect CKD. Since the number of people that can be affected are large it is highly unlikely for kidney imaging to be used for everyone in order to detect CKD. Therefore in this study CGAN is proposed for detecting CKD

Doctors inherently use some attributes and their inter-relationships from reports like that of blood reports and urine reports to conclude on diseases and disorders. If the attributes that contribute to CKD are identified clearly then it will lead to easier and effective identification of CKD early. Medical studies suggests that the relationship between attributes such as serum creatinine, blood pressure, haemoglobin, and albumin is different for subjects with CKD and healthy subjects. Therefore, first, the covariance among those attributes for both non-CKD and CKD subjects has to be estimated.

2 PRELIMINARIES

2.1 Dataset Collection

Chronic kidney disease is predicted by using dataset collection[1], which can be obtained from blood and urine reports from the test.The data set contains 400 samples and this samples has both CKD data set and non CKD data set, each sample has 24 predictive variables or features (11 numerical variables and 13 categorical (nominal) variables) and a categorical response variable (class). Each class has two values, namely, ckd (sample with CKD) and not ckd (sample without CKD).

2.2 Data Processing

Each categorical variable was coded to facilitate the computational processing by using the classifier [2]. For the values of RBC and pc, normal was coded as 1, and abnormal was coded as 0. For the values of PCC and ba, the present was coded as 1, and not present was coded as 0. For the values of htn, dm, cad, pe, and ane, yes was coded as 1 and no was coded as 0. For the value of appet, good was coded as 1 and poor was coded as 0. Although the first data description defines three variables sg, al, and su as categorical types, the values of those three variables are still numeric based, thus these variables were treated as numeric variables. All the categorical variables were transformed into factors. Each sample was given an independent number that ranged from 1 to 400. There is a large number of missing values in the data set, In general, the patients might miss some measurements for various reasons before making a diagnosis. Thus, missing values will appear in the data when the diagnostic categories of samples are unknown, and a corresponding imputation method is needed. After encoding the specific variables, the missing values within the original CKD dataset were processed and filled initially.Generally missing values are encoded with blanks or any other placeholder and the training a model with a dataset that has a lot of missing values can drastically impact the machine learning, those missing values were eliminated from the dataset and given to the feature extraction process.

3 LITERATURE SURVEY

Chronic renal disorder CKD may be global ill health with an Increasing prevalence and high cost [1]. 10% of individuals worldwide are affected by CKD [2], annually there are millions of people die prematurely of problems associated with CKD (World Kidney day). However, consistent with the NHS website there are not any signs or symptoms of CKD within the early stage. during this research, the info is used and obtained from [8]. Then, we perform k-Nearest Neighbour (KNN), Random Forest. The dataset used contains 24 attributes, containing 11 with real values and 14 with nominal values. Moreover, the info involves 400 instances. Their experiment result achieves 0.993 for the accuracy on F1- measure with 0.1084 RMS. an equivalent dataset is going to be utilized in this paper. Most of the presented research is used the WEKA data processing tool to research CKD data.

4 FEATURE EXTRACTION

The feature extraction process used is Recursive Feature Elimination, or RFE as popular feature algorithm, as it is easy to configure and use also it is effective in selecting the features in a training dataset that are more or most relevant in predicting the target variable. There are two important configuration options while using RFE they are chosen in the number of features to select and the choice of the algorithm used to help choose features. RFE works by searching for a subset of feature selection algorithm by starting with all features in the training dataset and also it removes the feature until the desired number remains this also improves the performance by ranking features and recursively removing the unwanted data in the dataset.

5 METHODOLOGY

The work proposed here uses classification techniques to predict the presence of chronic kidney disease in humans. The classifier used are KNN and CGAN classifier, here the KNN classifier was used as a reference to compare the performance of the CGAN classifier. The data set for chronic kidney disease was gathered and applied to the CGAN classifier to predict the disease and the performance of the classifier is evaluated based on accuracy, precision, recall, and F measure.

The working of the architecture is as follows: The dataset for CKD patients has been collected and fed into the classifier named KNN and CGAN. The prediction of CKD will be executed with the help of a tool known as Python. The dataset consists of attributes and values with class CKD and non-CKD. This tool will result in the accuracy that what percentage of patients are having CKD during a particular time. To improve the rate of prediction, a comparison of the two classifiers is done based on evaluation parameters. The experimental result is retrieved, which shows the best classifier between the two.

5.1 Evaluation Parameters

Sensitivity: Sensitivity is also known as True Positive Rate. It is used for measuring the percentage of unwell people from the dataset.

Sensitivity = Number of true positives/Number of true positives + Number of false negatives

Specificity is also known as True Negative Rate. It measures the percentage of healthy people that are exactly recognized from the dataset.

Specificity = Number of true negatives/Number of true negatives + Number of false positives

Precision and Recall: Precision is also known as positive predictive value. It is defined as the average probability of relevant retrieval.

Precision = Number of true positives/Number of true positives + False positives

Recall

Recall is defined as the average probability of complete retrieval. Recall= True positives/True positives + False negative

Accuracy: Accuracy is defined in terms of correctly classified instances divided by the total number of instances present in the dataset. Accuracy=Number of correctly classified samples/Total number of samples

5.2 K-nearest neighbor Classification

The K-Nearest Neighbor algorithm (K-NN) may be a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training example value in the feature space. K-NN is a type of instance-based learning. In K- NN Classification, the output may be a class membership. Classification is done by a majority vote of neighbors. If K = 1, then the class is a single nearest neighbor. In a common weighting scheme, an individual neighbor is assigned to a weight of 1/d if d is that the distance to the neighbor. The shortest distance between any two neighbors is always a straight line and the distance is known as Euclidean distance[4]. The limitation of the K-NN algorithm is it's sensitive to the local configuration of the info. The process of transforming the input data into a set of features is known as Feature extraction. In Feature space, extraction is taken place on data before applying he K-NN algorithm then we found error value of k by giving value to the k factor which is mentioned in the figure 6.1.

Evaluation

For evaluation the following algorithms are run:

Here *l* = number of clusters, *tp* = true positive, *fp* = false positive, *fn*

= false negative, *tn* = true negative.

Average Accuracy = $\frac{\sum_{i=1}^{l} \frac{tp_{i+}tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$

F-Measure = $\frac{(\beta^2 + 1)\text{precision}_M \text{recall}_M}{\beta \text{precision}_M \text{recall}_M}$

Precision = $\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$

5.3 Proposed method

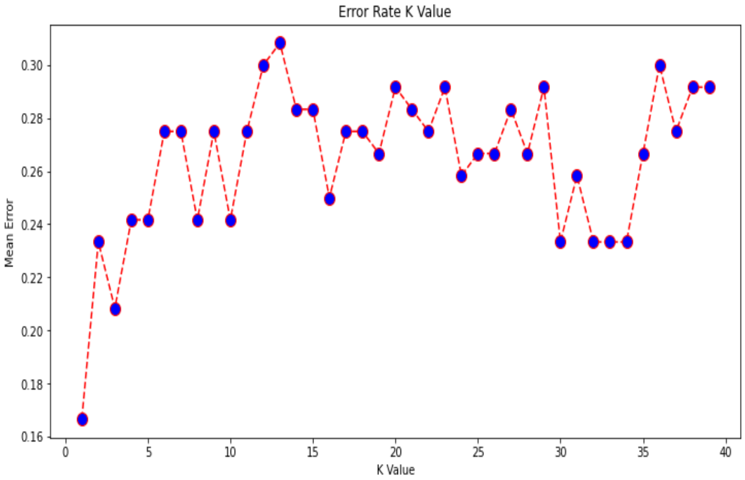
In this project, we have taken CGAN to get better accuracy than the KNN we know that the CGAN stands for conditional generative adversarial network whose generator and discriminator are conditioned during training by the information given to it. Here we are training our dataset, during training, the generator learns to produce realistic examples for each label in the training datasets, and the discriminator learns to distinguish fake example-label pairs from the real example label pairs. By using this concept in this project we are giving some training datasets to the CGAN and those results are used to predict the CKD and non-CKD patients during this process we got better performance than the KNN classifier which is shown in Table 1.

The below figures Fig 6.2 shows that CGAN performed better in terms of accuracy, precision, and f measure over different datasets, whereas KNN shows good results in calculating recall value. Thus we can say that CGAN performed better than KNN in the prediction of CKD in our analysis**.**

6 Experimental Result

This work is performed in a python tool, developed by Guido van Rossum. Python allows us to write and test the codes and it also allows us to debugging programs and integrating application with third-party web services. The experiment shows that the CGAN gives better accuracy than KNN but it has some loss in it by reducing the loss we can increase the accuracy.Fig 6.1 shows the error rate value of K and Fig 6.2 shows the accuracy level of the training set and the testset then the Fig 6.3 shows the Loss occured in training and testsets respectively.

Figure 6.1 Error rate value of K



Figue 6.2 Accuracy level

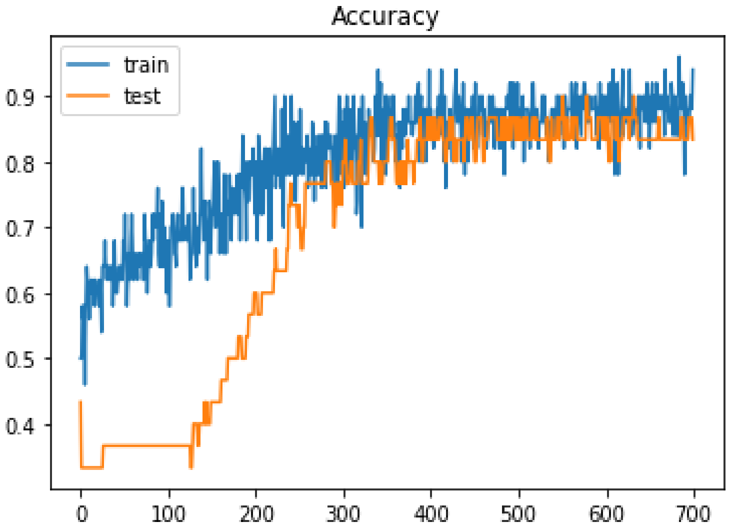


Figure 6.3 Loss level

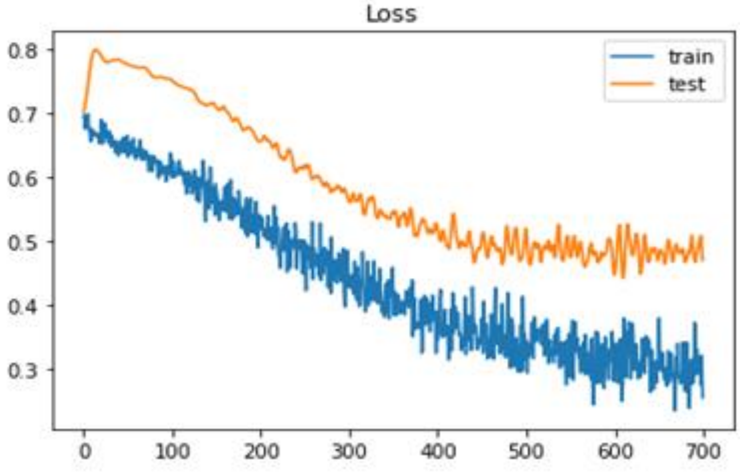


Table 1. Result Analysis

Name of Classifier	Evaluation Parameter			
	Accuracy	*Precision*	*Recall*	*F-measure*
KNN	0.76	0.86	0.69	0.79
CGAN	0.85	0.93	0.51	0.87

7 CONCLUSION

In this paper, a new decision support system is implemented for the prediction of CKD. In this paper, Chronic Kidney Disease is predicted using two different classifiers and a comparative study of their performance is done. From the analysis, we found that out of two classifiers CGAN and KNN, the CGAN classifier performed better than the other. The rate of prediction of CKD is improved.

8 FUTURE WORK

Other possible evolutionary techniques that may be used to improve the performance of the proposed classifiers. In this paper, CGAN is applied to detect CKD. We can also evaluate and compare the performance of the used classifiers with other existing classifiers also. CKD early detection helps in timely treatment of the patients suffering from the disease and also to avoid the disease from getting worse. Early prediction of the disease and timely treatment are the need for medical sector. New classifiers can be used and their performance can be evaluated to find better solutions of the objective function and by reducing the loss in the datasets we can improve the performance level in future work.

REFERENCES

1. Khamparia A, Gupta D, Nhu NG, Khanna A, Shukla B, Tiwari P (2019) Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE Access 7(1):7717–7727

2. Q.-L. Zhang and D. Rothenbacher, “Prevalence of chronic kidney disease in population-based studies: systematic review,” BMC public health, vol. 8, no. 1, p. 117, 2008.

3. K. V. Kokilam and D. P. M. P. Latha, "A review on evolution of data mining techniques for protein sequence causing genetic disorder diseases", *Computational*

4. *Intelligence & Computing Research (ICCIC) 2012 IEEE International Conference on*, pp. 1-6, 2012, December

5. Khamparia A, Nhu NG, Pandey B, Gupta D, Rodrigues JJ, Khanna A, Tiwari P (2019) Investigating the Importance of Psychological and Environmental Factors for Improving Learner’s Performance Using Hidden Markov Model. IEEE Access 7:21559–21571

6. P.Swathi Baby, T. Panduranga Vital ,"Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.

7. Lakshmanaprabu SK, Shankar K, Khanna A, Gupta D, Rodrigues JJ, Pinheiro PR, De Albuquerque VHC (2018) Effective features to classify big data using social internet of things. IEEE access 6:24196–24204.

8. Rosso R et al (2010) Chronious: an open, ubiquitous and adaptive chronic disease management platform for COPD, CKD and renal insufficiency. 32nd Annu Int Conf IEEE EMBS 2010:6850–6853.

9. Varughese S, Abraham G (2018) Chronic kidney disease in India: A clarion call for change. Clin J Am Soc Nephrol 13(5):802–804.

10. Chetty SDS, Naganna KSV (2015) Role of Attributes Selection in Classification of Chronic Kidney Disease Patients. Comput. Commun. Secur. (ICCCS), 2015 Int. Conf. on. IEEE, pp. 1–6.

11. Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms" IJCBR , ISSN (online): 2229- 6166,Volume 6 Issue 2 March 2015.

12. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors", *Information & Communication Technologies (ICT) 2013 IEEE Conference on*, pp. 1227-1231, 2013, April.

13. Mahfuzah Mustafa, Mohd Nasir Taib et.al, "Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image" Journal of Computer Science & Computational Mathematics, Volume 2, Issue 4, April 2012,pp 17-22.

14. Ross KK Leung, Ying Wang et.al, "Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype- phenotype risk patterns in diabetic kidney disease: a prospective case– control cohort analysis" BMC Nephrology 2013, pp 1-9.

15. DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi,"Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis", Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.