LITERATURE SURVEY

Early Detection of Chronic Kidney Disease using Machine Learning Algorithm

Team ID: PNT2022TMID35483

Team Members:

Sai Krishnan S - 2019105565

Arun Depak K G - 2019105512

Adithyaa Jagannathan Sudhakar - 2019105001

Kaviyarasan K – 2019105021

1. Introduction

Chronic kidney disease (CKD) has received much attention due to its high mortality rate. Chronic diseases have become a concern threatening developing countries, according to the World Health Organization (WHO). CKD is a kidney disorder treatable in its early stages, but it causes kidney failure in its later stages of life. In 2016, chronic kidney disease caused the death of 753 million people worldwide, where the number of males died was 336 million, while the number of females died was 417 million. It is called “chronic” disease because the kidney disease begins gradually and lasts for a long time, which affects the functioning of the urinary system. The accumulation of waste products in the blood leads to the emergence of other health problems, which are associated with several symptoms such as high and low blood pressure, diabetes, nerve damage, and bone problems, which lead to cardiovascular disease. Risk factors for CKD patients include diabetes, blood pressure, and cardiovascular disease (CVD). CKD patients suffer from side effects, especially in the late stages, which damage the nervous and immune system. Since CKD doesn’t show any symptoms at the initial stages, we cannot rule out that a person doesn’t have this particular disease without proper testing. Medical experts determine kidney disease through glomerular filtration rate (GFR), which describes kidney function. GFR is based on information such as age, blood test, gender, and other factors suffered by the patient. Regarding the GFR value, doctors can classify CKD into five stages. Table 1 shows the different stages of kidney disease development with GFR levels.

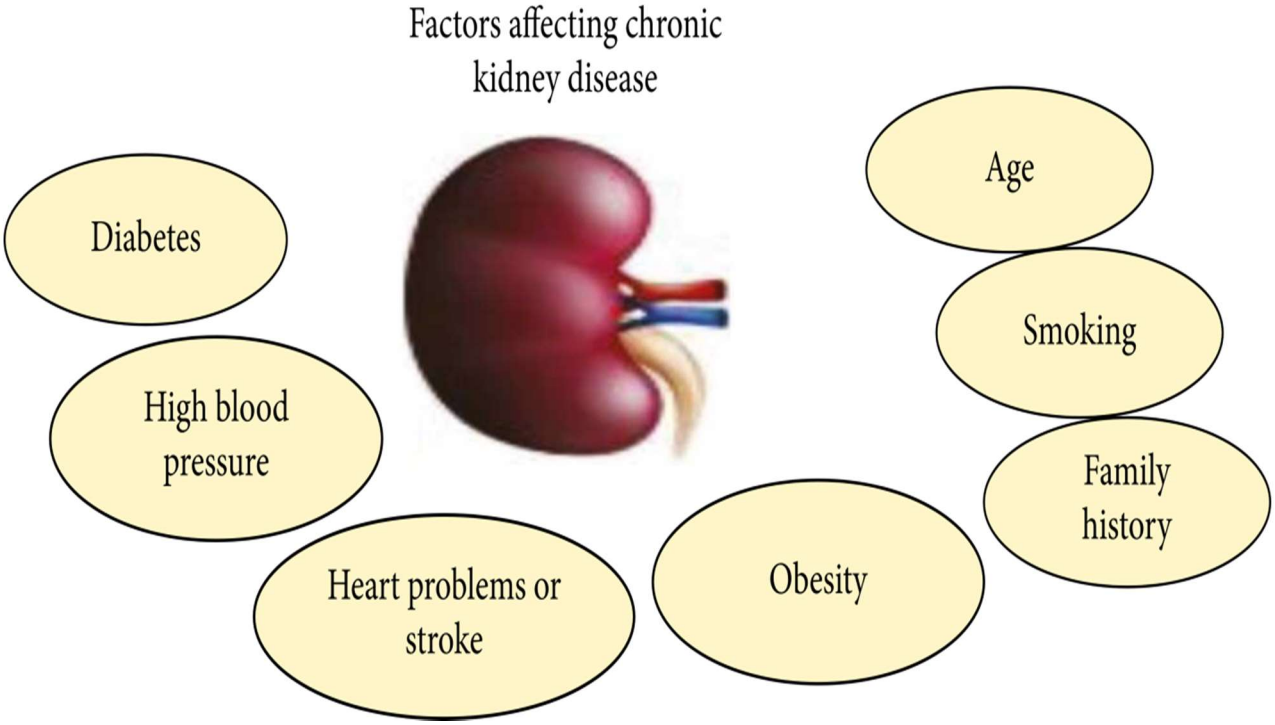
Table 1

The stages of development of CKD

Stage	Description	Glomerular filtration rate (GFR) (mL/min/1.73 m^2)	Treatment stage
1	Kidney function is normal	≥90	Observation, blood pressure control
2	Kidney damage is mild	60–89	Observation, blood pressure control and risk factors
3	Kidney damage is moderate	30–59	Observation, blood pressure control and risk factors
4	Kidney damage is severe	15–29	Planning for end-stage renal failure
5	Established kidney failure	≤ 15	Treatment choices

Early diagnosis and treatment of chronic kidney disease will prevent its progression to kidney failure. The best way to treat chronic kidney disease is to diagnose it in the early stages, but discovering it in its late stages will lead to kidney failure, which requires continuous dialysis or kidney transplantation to maintain a normal life. In the medical diagnosis of chronic kidney disease, two medical tests are used to detect CKD, which are by a blood test to check the glomerular filtrate or by a urine test to check albumin. Due to the increasing number of chronic kidney patients, the scarcity of specialist physicians, and the high costs of diagnosis and treatment, especially in developing countries, there is a need for computer-assisted diagnostics to help physicians and radiologists in supporting their diagnostic decisions. Artificial intelligence techniques have played a role in the health sector and medical image processing, where machine learning and deep learning techniques have been applied in the processes of disease prediction and disease diagnosis in the early stages. Artificial intelligence (ANN) approaches have played a basic role in the early diagnosis of CKD. Machine learning algorithms are used for the early diagnosis

of CKD. The ANN and SVM algorithms are among the most widely used technologies. These technologies have great advantages in diagnosing several fields, including medical diagnosis. The ANN algorithm works like human neurons, which can learn how to operate once properly trained, and its ability to generalize and solve future problems (test data). However, SVM algorithm depends on experience and examples to assign labels to the class. SVM algorithm basically separates the data by a line that achieves the maximum distance between the class data. Many factors affect kidney performance, which induce CKD, like diabetes, blood pressure, heart disease, some kind of food, and family history. The Figure 1 presents some factors affecting chronic kidney disease.



2. Materials and Methods

Figure 2 shows the general structure of CKD diagnosis. In pre-processing, mean method and mode method was used to compute the missing values. The features of importance associated with the CKD diagnosis were selected using Recursive Feature Elimination algorithm. These selected features were fed into classifiers for disease diagnosis. Various classifiers show promising results for diagnosing a dataset into CKD or a normal kidney.

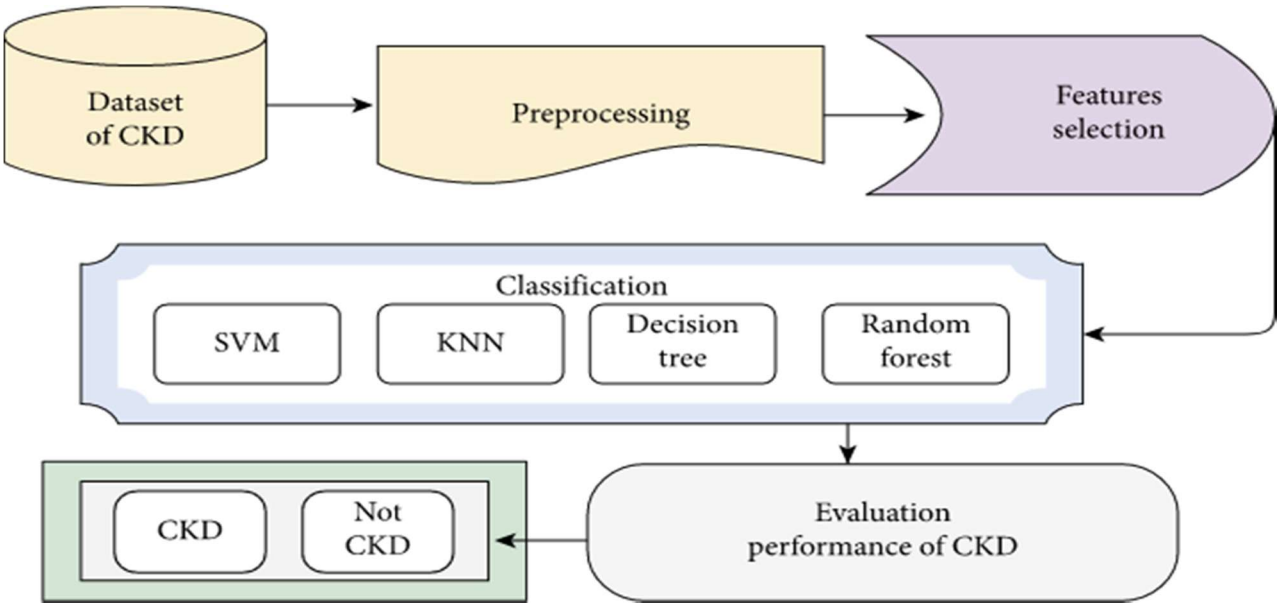


Fig 2: PROPOSED SYSTEM

2.1. Dataset

The dataset will be formed using several features divided into numeric features and categorical features, in addition to the class features, such as “CKD” and “NOTCKD” for classification. Some of the features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anaemia, etc. The diagnostic class contains two values: CKD and NOTCKD. All features contained missing values are filled using mean and mode values.

2.2. Pre-processing

The dataset contains outliers and noise, so it must be cleaned up in a pre-processing stage. The pre-processing stage includes estimating missing values and eliminating noise, such as outliers, normalization, and checking of unbalanced data. Some measurements may be missed when patients are undergoing tests, thereby causing missing values. The simplest method to handle missing values is to ignore the record, but it is inappropriate with small dataset. We can use algorithms to compute missing values instead of removing records. The missing values for numerical features can be computed through one of the statistical measures, such as mean, median, and standard deviation.

However, the missing values of nominal features can be computed using the mode method, in which the missing value is replaced by the most common value of the features.

2.3. Features Selection

After computing the missing values, identifying the important features having a strong and positive correlation with features of importance for disease diagnosis is required. Extracting the vector features eliminates useless features for prediction and those that are irrelevant, which prevents the construction of a robust diagnostic model. The Recursive Feature Elimination (RFE) algorithm is very popular due to its ease of use and configurations and its effectiveness in selecting features in training datasets relevant to predicting target variables and eliminating weak features. The RFE method is used to select the most significant features by finding high correlation between specific features and target (labels). It is noted that albumin feature has highest correction (17.99%), featured by 14.34%, then the packed cell volume feature by 12.91%, and the serum creatinine feature by 12.09%.

2.4. Classification

Data mining techniques can be used to define new and understandable patterns to construct classification templates. Supervised and unsupervised learning techniques require the construction of models based on prior analysis and are used in medical and clinical diagnostics for classification and regression. Four popular machine learning algorithms used are SVM, KNN, decision tree, and random forest, which give the best diagnostic results. Machine learning techniques work to build predictive/classification models through two stages: the Training Phase, in which a model is constructed from a set of training data with the expected outputs, and the Validation Stage, which estimates the quality of the trained models from the validation dataset without the expected output. All algorithms are supervised algorithms that are used to solve classification and regression problems.

2.5 Splitting Dataset

The dataset will be divided into 75% for training and 25% for testing and validation.

2.6 Evaluation Metrics

Evaluation metrics are used to evaluate the performance of the four classifiers. One of these measures is through the confusion matrix, from which the accuracy, precision, recall can be extracted by computing the correctly classified samples (TP and TN) and the incorrectly classified samples (FP and FN), as shown in the following equations:

Accuracy = (TN+TP) / (TN+TP+FN+FP) * 100%
Precision = TP/(TP+FP) * 100%
Recall = TP/(TP+FN) * 100%

Where, TN is True Negative, TP is True Positive, FN is False Negative, and FP is False Positive.

3. Conclusion

This study is done to provide insight into the diagnosis of CKD patients to tackle their condition and receive treatment in the early stages of the disease. The dataset will be processed to remove outliers and replace missing numerical and nominal values using mean and mode statistical measures, respectively. The RFE algorithm can be applied to select the most strongly representative features of CKD. Selected features are fed into classification algorithms: SVM, KNN, decision tree, and random forest. The parameters of all classifiers will be tuned to perform the best classification, so that all algorithms can reach promising results.

References

- Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods Yedilkhan Amirgaliyev; Shahriar Shamiluulu; Azamat Serek2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT) Year: 2018 | Conference Paper | Publisher: IEEE

- A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction; Linta Antony; Sami Azam; Eva Ignatious; Ryana Quadir; Abhijith Reddy Beeravolu; Mirjam Jonkman; Friso De Boer IEEE Access

Year: 2021 | Volume: 9 | Journal Article | Publisher: IEEE Cited by: Papers (3)

- de Almeida et al. in their work applied Decision tree, Random Forest, Support Vector Machine (SVM) and also used SVM with linear, polynomial, sigmoid and RBF functions. For their research, they used the MIMIC-II database. They concluded that random forest and Decision tree got the best result in the form of prediction accuracy of 80% and 87% respectively.

- Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Drall et al. worked on CKD dataset given by UCI with 400 instances and 25 attributes. Firstly, data was pre-processed, the missing data was found, filled with 0, then transformed and applied on the dataset. After pre-processing, authors applied algorithm for important attributes and found 5 most important features and then the classification algorithm: Naïve Bayes and K-Nearest Neighbour. The gotten result KNN achieved the highest accuracy.

- Abrar, Tahmid, Samiha Tasnim, and Md Hossain. Early detection of chronic kidney disease using machine learning. Diss. Brac University, 2019.

- Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P., Aljaaf, A.J. (2020). Performance-Based Prediction of Chronic Kidney Disease Using Machine Learning for High-Risk Cardiovascular Disease Patients,2020.

- Makino, Masaki, et al. "Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning." Scientific reports 9.1 (2019): 1-9.