Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/16701
DOI URL: http://dx.doi.org/10.21474/IJAR01/16701

RESEARCH ARTICLE

AN EFFECTIVE MACHINE LEARNING APPRAOCH FOR CHRONIC KIDNEY DISEASE DETECTION

G. Nagarjuna Reddy, B. Dhana Lakshmi, C. Jaya Sree, A. Lokesh and G. Madhuri
Department of Electronics & Communication Engineering, N.B.K.R. Institute of Science & Technology, Vidyanagar, Andhra Pradesh, India.

Manuscript Info

Manuscript History
Received: 15 February 2023
Final Accepted: 19 March 2023
Published: April 2023

Key words:-
Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbour, Naive Bayes Classifier, and CKD

Abstract

Chronic kidney disease (CKD) is a global health problem with high mortality and morbidity and mortality. Real-time performance using machine learning. In this study, we introduce machine learning for CKD diagnosis. CKD data is from the University of California, Irvine (UCI) Machine Learning Repository, which contains many missing values. KNN assignment selects multiple completed models with the best values to predict missing data for each incomplete model and is used to load missing values.Although patients may ignore certain measures for a variety of reasons, missing data is often found in real clinical settings. After solving the missing data, models are constructed using machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbor, Naive Bayesian classifier, and feedforward neural network). Random forest machine learning models are the most accurate in this task.

Introduction:-

A global public health issue, chronic kidney disease (CKD) affects 10% of the world's population [1]. This illness is characterised by a gradual decline in renal function that ultimately results in a total loss of renal function. Early on, CKD does not have noticeable symptoms. Because of this, the illness might not be discovered until the kidney has lost around 25% of its functionality.It can cause cardiovascular disease to develop. CKD is a pathologic illness that progresses and cannot be reversed. Thus, it is very important to detect and diagnose CKD in its early stages so that patients can start therapy right away to slow the disease's progression. A computer programmer that calculates and extrapolates task-related information and determines the traits of the relevant pattern is referred to as machine learning.This technology makes it possible to identify diseases accurately and affordably; as a result, it may be a promising technique for diagnosing CKD. With the advancement of information technology, it has evolved into a new kind of medical instrument and has a wide range of potential applications.Algorithms based on regression, trees, probabilities, decision surfaces, and neural networks were frequently successful in these models.This study compares the effectiveness of various machine learning algorithms for the prediction of chronic renal disease. This papergives a summary of the available research on this subject, point out any gaps in the knowledge, and suggest a way for assessing the correctness of various algorithms. The findings of this study will aid in determining the best machine learning algorithm for identifying websites with chronic kidney disease and offer suggestions for enhancing the precision of such detection methods.

Corresponding Author:- G. Nagarjuna Reddy
Address:- Department of Electronics & Communication Engineering, N.B.K.R. Institute of Science & Technology, Vidyanagar, Andhra Pradesh, India.

Literature Survey:-

L. Zhang et al., [2] used a cross-sectional survey of a nationally representative sample of Chinese adults.The prevalence of chronic kidney disease was high in developing countries. However, no national survey of chronic kidney disease had been done incorporating both estimated glomerular filtration rate (eGFR) and albuminuria in a developing country with the economic diversity of China. This aimed to measure the prevalence of chronic kidney disease in China with such a survey.

A. Singh et al., [3]proposed a methodology for building predictive models using temporal data in electronic health records (EHRs) can have a significant impact on managing chronic diseases. In this paper, the authors evaluated three different approaches that used machine learning to build predictive models using temporal EHR data of a patient. The study presented three methods, one that did not use temporal information and two methods that captured temporal information. By addressing challenges related to irregularly sampled data and varying lengths of patient history, these methods could help improve the accuracy of predictive models and ultimately lead to better management of chronic diseases.

Machine learning techniques have been increasingly used in the field of medical diagnosis due to their high accuracy rates in classification. **H. Polat et al., [4]**proposed a methodology aims to detect chronic kidney disease(CKD) at an early stage to improve management and reduce mortality rates. The study utilized wrapper and filter methods on a CKD dataset. For the filter approach, two evaluators were used - CfsSubsetEval with Greedy stepwise search engine and Filter SubsetEval with Best First search engine. These methods address the challenges of feature selection and potentially lead to better diagnosis and treatment of CKD. Overall, the study demonstrates the potential of machine learning in improving healthcare outcomes**.**

S. Ramya et al.,[5] proposedan approach for reducing diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD according to its gravity. By analysing different algorithms like Basic Propagation Neural Network, RBF and RF. The analysis results indicates that RBF algorithm gives better results than the other classifiers and produces 85.3% accuracy.

C. Barbieri et al., [6]presented a work about Chronic Kidney Disease (CKD) anaemia is one of the main common co morbidities in patients undergoing End Stage Renal Disease (ESRD). Iron supplement and especially Erythropoietin Stimulating Agents (ESA) have become the treatment of choice for that anaemia. However, it is very complicated to find an adequate treatment for every patient in each particular situation since dosage guidelines are based on average behaviours, and thus, they do not take into account the particular response to those drugs by different patients, although that response may vary enormously from one patient to another and even for the same patient in different stages of the anaemia.

V.papademetrion et al., [7]surveyed early stages of chronic kidney disease are associated with an increased cardiovascular risk in patients with established type 2 diabetes and macro vascular disease. The role of early stages of chronic kidney disease on macro vascular outcomes in prediabetes and early type two diabetes mellitus is not known. In high-risk patients with deglycation (pre-diabetes and early diabetes), mild and moderate chronic kidney disease significantly increased cardiovascular events.

Methodology:-

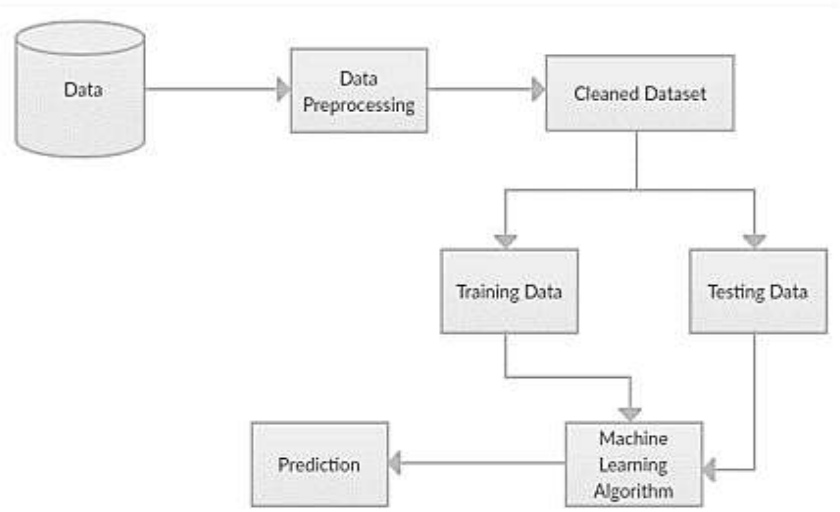


Fig 1:- Flowchart of proposed approach

Dataset

Fig 1illustratesachronic kidney disease detection approach and some datasets have generated from various websites including Kaggle [8] for this study. The chronickidney disease dataset in the University of California, Irvine (UCI) repository.The dataset has 400 samples from two different sources. There are a total of 25 features, including 11 numeric features, 13 nominal features, and 1 class feature. There are several missing values in the dataset. Here, age, blood pressure, gravity, albumin, sugar, red blood cells etc.

Pre-Processing

Many data points may be irrelevant or missing. This is controlled by previous data as shown in Fig 1. Data pre-processing is the process of transforming raw data into useful and useful data. It includes dealing with missing data, noisy data, and other issues. There are missing values in all features except diagnosis.Data were not equal or unequal, with 250 patients (62.5%) with 'chronic kidney disease' and 150 patients (37.5%) with 'NOT Chronic kidney disease'. The K-Nearest Neighbor Imputer method will be used to replace missing and nominal data. Use the 'map' function to convert categorical variables to numeric values.

Training and Testing using dataset:

The train-test separation is a way to determine or evaluate the effectiveness of a machine learning [9-11] algorithms. Using this technique, the data is splitted into two groups. The training data is used to train the machine learning model. This data is used to evaluate machine learning models. The purpose of this process is to calculate the performance of machine learning models on new data that has not been used to train before. The dataset will be split into 30% for testing and 70% for training.

Classification:-

The paper involves machine learning and focuses on monitoring two key elements - distribution and recovery - in a classification problem where ideas are categorized as either normal or abnormal. To achieve this, several classification models such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, Naive Bayes, and Neural Network Classifier were selected for analysis.

Logistic Regression:

Logistic regression is a method used to estimate the probability of a binary event based on one or more variables. This means that logistic regression can be used to predict outcomes with two possible outcomes, such as pass or fail, yes or no, 0 or 1. Unlike other regression models, logistic regression is the analysis of data and the relationship between a variable and one or more independent variables, which can be nominal, rank, range, or rank ratio. Instead of a linear function, the algorithm uses a complex value function called a "sigmoid function" or "logistical function" and assumes that it is limited to function values in the range 0 to 1 as shown in Eq.(1).

$$0 \leq h_\theta(x) \leq 1$$

a) Support Vector Classifier:

Another good technique in machine learning is the support vector machine [12]. Each data is represented as a point in n-dimensional space in the support vector machine method, and the support vector machine algorithm creates a dividing line for the two classes; this dividing line is often called a hyperplane. A support vector machine looks for the nearest points called support vectors and creates a line connecting them. SVM then creates a divider perpendicular to the connected line and splits it in half. Margins should be as wide as possible to properly categorize information. In this case, the margin is the distance between the hyperplane and the vectors. In practice it is impossible to separate complex nonlinear data; so, support vector machines use kernel methods to convert low dimensional fields to high dimensional ones.

b) Random Forest:

In supervised machine learning, the Random Forest method is used to address classification and regression issues. Multiple decision trees are created in the random forest and all decision trees are combined to provide a reasonable estimate. Random forest is an integrated method that provides suitable models by combining baseline models. A random forest is a distribution that contains decision trees for many combinations of given data and is a way to improve the accuracy of estimating these data. The random forest takes predictions from all decision trees and combines the results with a majority vote. The number of trees is determined not only for accuracy, but also to prevent overfitting of the model. Random forest works in two phases; first, by combining N decision trees to generate a random forest, and second, by predicting each tree created in the first step. The random forest algorithm can be performed with the following steps

Step-1: First select K points from the training.
Step 2: Build a decision tree from the data points.
Step 3: By choosing the number N for the decision tree we want to create.
Step 4: Repeat steps 1 and 2 to build the decision tree.
Step 5: Find the prediction of each decision tree for the new data.

K-Nearest Neighbor:

One of the most efficient classification algorithms in Machine Learning is K-Nearest Neighbor. This algorithm belongs to the supervised machine learning and is immensely used in pattern recognition, data mining and intrusion detection. This algorithm can be used to solve both the classification and regression problems. Imaginary boundary will be created to classify the data. This will be done based on the idea of similarity (sometimes called distance, proximity or closeness) with some mathematics like calculating the distance between the points on a graph. The popular method used for calculating the distance is Euclidian distance. The mathematical formula is shown in Eq. (2).

$$\begin{aligned} d(p,q) &= d(q,p) \\ &= \sqrt{(q1-p1)^2 + (q2-p2)^2 + \cdots .. (qn-pn)^2} \\ &= \sqrt{\sum_{i=1}^{n} (qi-pi)^2} \end{aligned} \qquad \text{Eq. (2)}$$

KNN executes the above given formula to find the intermediate distance between every single data point and the testing data. Then it computes the possibility of points which are identical to the testing data and classifies them based on what all points give the best probability. While using KNN for classification, the outcome is computed from the class which has the best frequency from the K closest resembling instances. Class probabilities are computed using the standardized prevalence of cases that is a component to each of the class in the set of 'K' closest resembling cases for every new data instance.

Naive Bayes:

Naive Bayes this approach is simple and probabilistic.In this way, a table will appear which is the model and is modified by the training data. A "probability table" is based on the main feature of the class that it should look at to predict new observations. It holds continuous and discrete data. It is insensitive to irrelevance properties.Although predictive, it requires more memory than SVM or simple logistic regression. It contains many things, especially for models with many different features. Naive Bayes can be used in applications such as recommendations and predictions about cancer or post-radiation growth.

Neural Network:
Neural networks are a family of algorithms designed to identify relationships in a dataset through a process that mimics the way the human brain works. In this sense, a neural network refers to a system of neurons, whether organic or artificial. Neural networks can adapt to changing inputs; so, the network can produce good results without having to recreate the output model. The concept of neural networks started from artificial intelligence and is rapidly gaining popularity in the development of trading systems. Neural networks work similarly to the neural networks of the human brain.A "neuron" in a neural network is a mathematical function that collects and distributes information according to a pattern. This network is similar to statistical methods such as curve fitting and regression analysis.

Results and Discussions:-

The results of all the machine learning algorithms were compared and analysed with the help of various performance metrics.The dataset is randomly divided into 70% for training and 30% for testing and validation. The Recursive Feature Elimination method was presented to select the irrelevant subset features. Then, the select features were processed by employing classifiers for diagnosis of CKD.

While the proposed system has obtained accuracy of 99% with Random Forest method. Finally, it is observed that the proposed has optimal results compared with existing systems.Twenty-four numerical and nominal features were introduced from 400 patients with CKD. Due to the neglect of some tests for some patients, some computation methods were applied to solve this problem. To solve the missing numerical values, mean method was used for missing nominal values, the mode method was used.

In addition to building and training the machine learning models, we also developed a Flask web application to make the chronic kidney disease detection system more accessible to users. The web application enables users to upload dataset and receive a prediction of whether the chronic kidney disease or not chronic kidney disease based on the trained models.

Here Fig 2displays the home page of A Machine Learning Methodology for Diagnosing Chronic Kidney Disease using web application.

Fig 2:- Home page for chronic kidney disease detection.

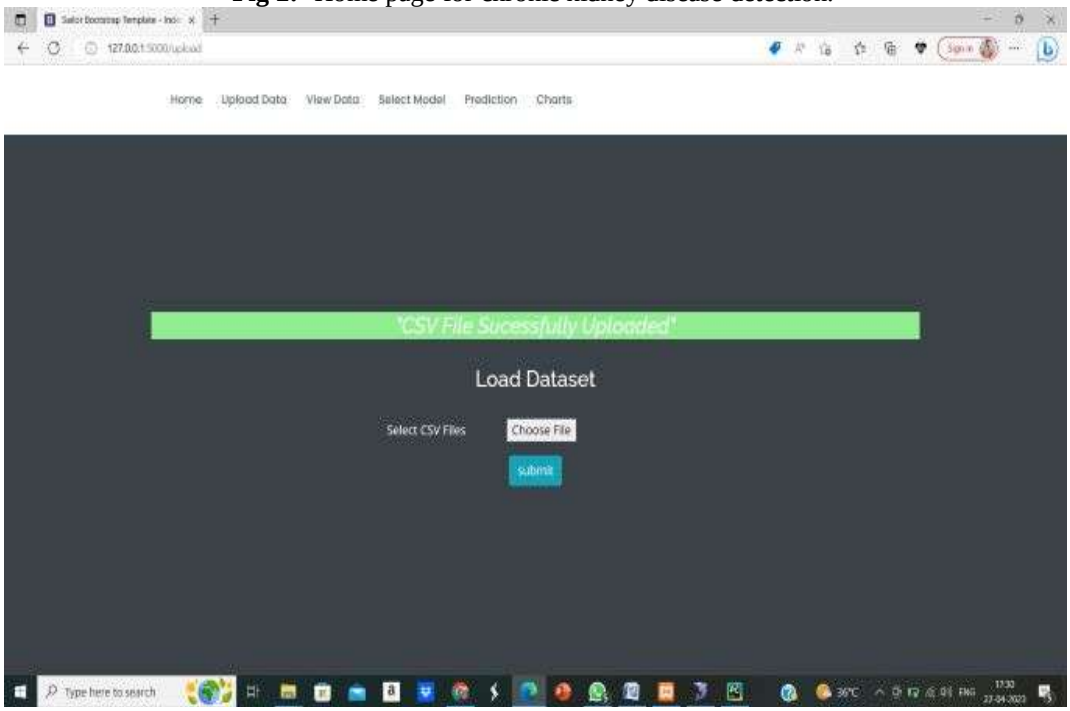
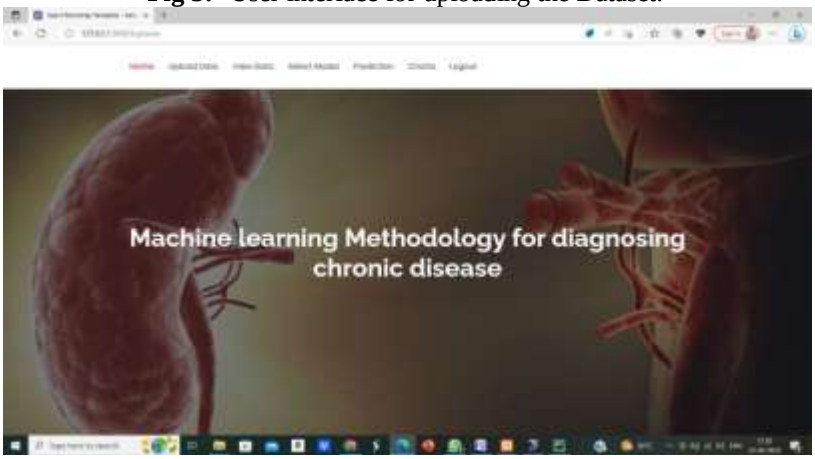


Fig 3:- User interface for uploading the Dataset.



Then from Fig 3 upload the Dataset on to the software. Then by clicking on the predict button, the user software will detect the chronic kidney disease by using the trained data and the test output will be displayed with the input values down with the chronic kidney disease.

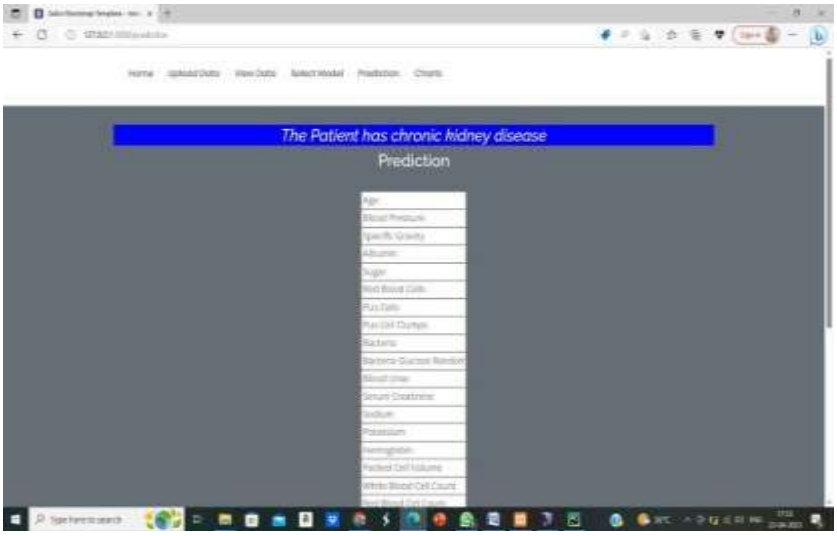


Fig 4:- User Interface of with test output for chronic kidney Disease prediction.

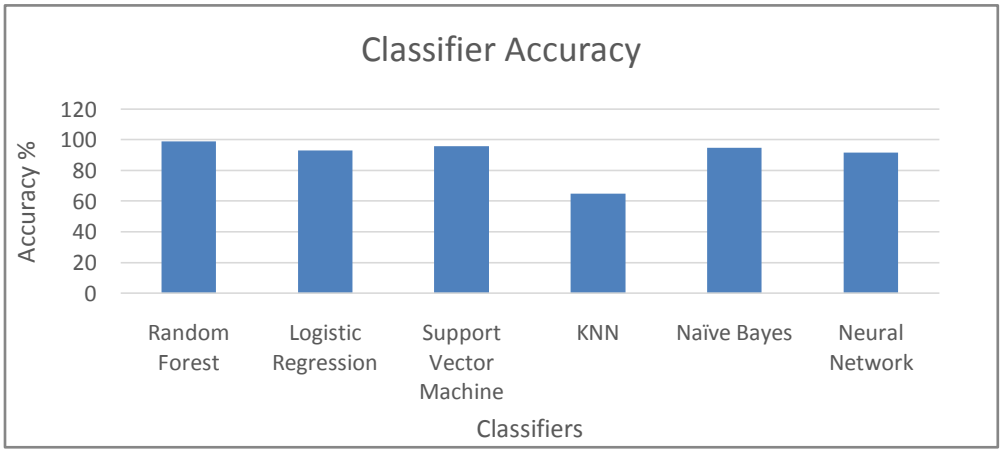


Fig 5:- Accuracy of various classifiers.

This paper focuses on five popular algorithmsfor classification: Logistic regression, neural network, KNN, Support vector machine, Naive Bayes, and Random Forest. All these approaches are based on supervised learning. We are determining the best method considering Accuracy. The Fig 5 depicts that Random Forest was the best method to use to find chronic kidney disease.

Accuracy shown in Eq. (3) is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. (3)}$$

S.NO	ALGORITHM	ACCURACY
1	Random Forest	99.17%
2	Logistic regression	93.33%
3	Support vector machine	95.83%
4	K-Nearest Neighbor	65.0%
5	Naive Bayes	95.0%
6	Neural Network	91.66%

Table 1:- Prediction of classical algorithms.

According to Table1,supervised learning provides the most accurate predictions compared to unsupervised learning. In supervised learning, Random Forest provides the most accurate predictions compared to support vector machine and K-means. That's why we choose Random Forest to provide the most accurate predictions. From this estimate, we can make some suggestions for improving chronic kidney disease. Based on these Algorithms, 99% accuracy is attained which is 25% better thanKNN. By using SVM and Naïve Bayes, a 95% accuracy is obtained which is 2% better than Logistic Regression. By using Neural Network, 91% accuracy is achieved which is far better than KNN which has only an 65% accuracy.

Conclusion:-

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. We speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. The performance of the models is evaluated based on an Accuracy. The results of the research showed that Random Forest model better predicts CKD in comparison to the other models taking Accuracy under consideration. This system would help detect the chances of a person having CKD further on in his life which would be really helpful and cost-effective people. This model could be integrated with normal blood report generation, which could automatically flag out if there is a person at risk. Patients would not have to go to a doctor unless they are flagged by the algorithms. This would make it cheaper and easier for the modern busy person

References:-

1. Z. Chen et al., “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” Chemometer. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.
2. L. Zhang et al., “Prevalence of chronic kidney disease in china: a cross-sectional survey,” Lancet, vol. 379, pp. 815-822, Aug. 2012.
3. A. Singh et al., “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” J. Biomed. Inform. vol. 53, pp. 220-228, Feb. 2015.
4. H. Polat, H.D. Mehr, A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” J. Med. Syst., vol. 41, no. 4, Apr. 2017.
5. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

6. C. Barbieri et al., “A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis,” Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.
7. V. Papademetriou et al., “Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study,” Am. J. Med., vol. 130, no. 12, Dec. 2017.
8. www.kaggle.com
9. N. R. Hill et al., “Global prevalence of chronic kidney disease – A systematic review and meta-analysis,” Plos One, vol. 11, no. 7, Jul. 2016.
10. M. M. Hossain et al., “Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts,” IEEE Trans. Ultrason. Ferr. vol. 66, no. 3, pp. 551-562, Mar. 2019.
11. M. Alloghani et al., “Applications of machine learning techniques for software engineering learning and early prediction of students’ performance,” in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246-258.
12. G Nagarjuna Reddy, K Nagi Reddy, “A Robust Machine Learning Approach for Multiclass Alzheimer’s Disease Detection using 3D Brain Magnetic Resonance Images”, Journal of Engineering Research, Vol. 10, Issue 2(A), pp:82-94, DOI: https://doi.org/10.36909/jer.10511.