Research Article

Machine Learning Algorithms for Chronic Kidney Disease Risk Prediction

S. Jaya priya[1], A.Thamaraiselvi[2], and S.Sinduja[3]

[1]PG Scholar, Department of Computer Science & Engineering, Vivekanandha College of Engineering for Women(Autonomous), Tiruchengode - 637205.
Email: priyacsejaya@gmail.com
[2]Assistant Professor, Department of Computer Science & Engineering, Vivekanandha College of Engineering for Women(Autonomous), Tiruchengode - 637205.
Email: thamaraiselvi@vcew.ac.in
[3]Assistant Professor, Department of Computer Science & Engineering, Vivekanandha College of Engineering for Women(Autonomous), Tiruchengode - 637205.
Email: sindujacse@vcew.ac.in

Article History: Received: 5 April 2021; Accepted: 14 May 2021; Published online: 22 June 2021

ABSTRACT:

In today's world, everyone tries to be health-conscious, but owing to work and a hectic schedule, people only pay attention to their health when signs of sickness appear. Chronic Kidney Condition is a disease that does not have any symptoms or, in some situations, does not have any disease-specific signs. As a result, it is difficult to forecast, identify, and prevent such a sickness manually, which could result in lasting health damage. Machine learning, which excels at prediction and analysis, provides a ray of hope in this dilemma. We studied CKD patient data and presented a system for predicting CKD risk using machine learning algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbor (K-NN). We used data from 455 patients. Here, an online data set from the UCI Machine Learning Repository and a real-time dataset from Khulna City Medical College are employed. For the development of our system, we used Python as a high-level interpreted programming language. We used a 10-fold CV to train the data using a Hybrid ensemble technique. The hybrid ensemble technique achieves 97.12 % accuracy, whereas ANN achieves 94.5 %. This technology will aid in the early detection of chronic kidney disorders..

Keywords: chronic kidney disease, hybrid ensemble technique, risk prediction, machine learning algorithm.

1 Introduction

If class labels are uniformly dispersed, machine learning techniques for data classification can very effectively be used for classification precision. However, in case of classifying the unequalled data which differ in class labelling, these typical algorithms have less or less learning performance. One or more algorithms can be coupled and have the reasonable accuracy to ensure the accuracies of prediction and classification systems. Assembly is defined the process of joining multiple algorithms. The research now underway predicts the probability of renal disease predictions from patient data. CKD has recently been a leading cause of death due to the change of citizens' normal way of life.

Machine Learning is an artificial intelligence branch that aims to provide computer methods to accumulate change and update intelligent systems knowledge. Artificial intelligence (AI) allows systems to observe from environments, execute certain features and increase the likelihood of success in fixing real world challenges. AI turns out to be an interesting field with technological improvements and scientific growth. It therefore leads to a growing attention on ML techniques. Machine learning (ML) is a significant method for data analysis that iteratively learns from the available data with the aid of learning algorithms.

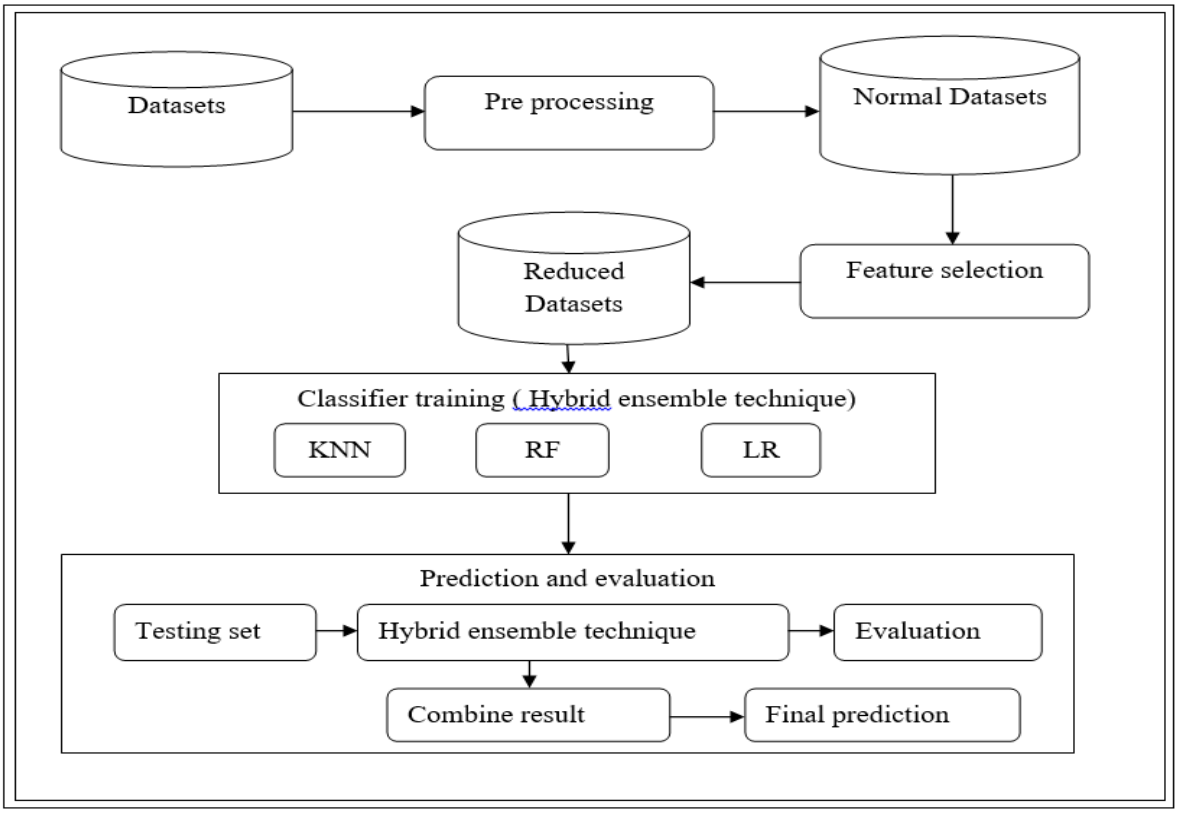

Figure 1: Hybrid ensemble technique

We have a model defined to certain parameters and learning involves running a computer program that optimizes the parameters of the model with the use of training data or past knowledge. The model can anticipate how future or descriptive forecasting can be made to gather information from data or both" Machine learning is an app for artificial intelligence that allows systems to automatically learn and improve without being explicitly coded. The objective of machine training is the development of programmes, which can use and learn the data by themselves.

2 Method package:

The whole procedure is a predictor or classifier accuracy strategy. The ensemble approach uses a combination of models to generate an enhanced composite model for performance improvements. The primary idea behind the technique of the ensemble is to group several "weak learners" into a "strong learner." Bagging and boosting are two typical strategies for ensembles. Boosting and bagging are both predictable and classifiable. Bagging is a strategy used by the ensemble, with numerous independent predictors and students and their outcomes pooled by majority voting, while the predictors or students are sequentially not independent by boosting them. This sequence approach is because each classification "takes greater attention" to the training times that were incorrectly classified by the preceding classification by giving weights to each classification.

Objective of the paper:

- To detect CKD in accurate way by implementing Hybrid ensemble technique this achieves maximum accuracy.
- The inputted dataset is trained by hybrid ensemble technique hence the accuracy of prediction attains high accuracy.
- Large amount of dataset is utilized for training which leads to maximum accurate prediction.

3 Literature Survey

There are many people worldwide who suffer from chronic renal disease, including Bidri Deepika et.al (2020). Many people suddenly develop diseases because of several risk factors like food, environment and livelihood. Invasive, costly, time consuming and potentially risky diagnoses of chronic renal disease are usually used. Therefore, numerous individuals without therapy reach a late stage, particularly in nations that have limited resources. The early detection approach of the disease therefore remains vital, particularly in underdeveloped nations, where diseases are generally subsequently detected. The work was strongly driven by finding a solution to the aforesaid concerns and avoiding any downsides. Chronic kidney disease (CKD), which leads in gradual loss of the function of the kidney, is a kind of kidney disease. Due to numerous patient living situations, this phenomena might be observed over the course of months or years. The aim is to construct an application for the early detection of CKD in real time with machine learning methods (Naive-Bayes and KNN-algorithms).

Reshma Set.al (2020), has conducted a chronic renal disease, defines a noncharacteristic kidney function or renal failure that spans over months or years. It is also known to occur in chronic renal disease. During screening, chronic kidney disease is usually discovered in patients who are known to be at risk from kidney disease (CKD) such as high blood pressure patients or diabetes. Therefore, early prediction to fight and cure the disease is crucial. The aim of this work is to employ machine learning techniques for the CKD classification, such as the Ant Colony Optimization (ACO) technique and the SVM classification. The final output will determine whether or not the person has CKD with minimal characteristics.

Sai Prasad Potharaju and M.Sreedevi (2016) The unbalanced data from data of a kind where the class ratio differs. In actual life, data analyzes are easily done. Most machine learning algorithms tend to harm the majority class in the event of imbalanced data, and so misunderstand the minority class. Therefore in this article the question of an unbalanced data classification problem is being discussed systematically through the application of rules-based ensemble training technology, such as bagging, stimulation, voting and stacking to create models. Initially, the data collected is unequaled. First, the uneven data is balanced using a sampling approach called SMOTE algorithm. Different learning approaches were then utilized to better forecast. The results suggest that the selected model template may effectively decrease the problem of imbalanced data misclassification. However, as the uneven rate of class increases, i.e. for big data, this template model cannot be properly classified.

Shubham Gupta (2019) often deals with huge amounts of data in the medical area. Conventional procedures can influence the outcomes while handling enormous data. Machine learning algorithms can be used to identify facts in medical research, especially in the prediction of diseases. Early disease recognition is vital for analyzing drugs and specialists in patients. Algorithms such as Decision Baumes, Vector Machine, Multilayer Perceptron, K-Nearest Neighbours. Classification procedures etc are used to determine different distresses. The use of algorithms to machine learning can lead to quick and accurate disease prediction. This study article discusses the methods used to predict various diseases and types of machine learning approaches. This work looked largely at the prediction of chronic renal diseases, machine learning, cardiomyopathy, diabetes and breast cancer. The research also looks at the hybrid method, which boosts individual classifications' performance.

Sahil Sharma (2018), in attendance Chronic renal disease is a common term for different diverse conditions that influence kidney structure and function. It's a high death rate disease. In this study, the authors suggested a very efficient two-stage hybrid ensemble technique. The potential of individual classification methods is merged in two phases of hybrid ensemble classification. In addition, 8 parameters of prime relevance were optimally picked from the 24 data set parameters of the study by the authors. The selected parameters (features) represent the intersection of the two sets; 1 contains vital parameters arranged for a decrease in diagnostic contribution and other set of parameters which are classified in decreasing order in their contribution to the Machine Learning process. Results from this group-classifying set for the ideally chosen reduced feature set (with 8 parameters) along with the whole feature set (with 24 parameters) are predictive (2-class) 100% accuracy, 1, 1 precision, and 1 F value.

Syrage Zeynu and Shruti Patil (2018), The syrup of the kidney has impacted the body, and can bring grave disease and death. The most important function in high-performance illness prediction and machine learning and data mining approaches are used to aid decision-makers to gather and comprehend information. Classification technique performance depends on the characteristics of the data set. Improving the accuracy of the

classification approach by reducing feature size and the ensemble being used, or by combining an algorithm model. In this research the strategies employed for the categorization of chronic kidney disease have been K-Nearest Neighbor, J48, the Artificial Network, Nave Bayes and Support for Vector Machines. Build two crucial models to prevent chronic kidney disease. Use the approach and model of feature selection. Using ranker search engine and wrapper subset evaluator with optimal first-engine information acquire attributes were employed in order to create chronic prognosis for renal diseases. The result has shown that 99% of the K nearest nearest neighbor selected by the Wrapper Sub-Set Evaluator is 99% accurate.

4 Proposed Methodology

Chronic Kidney Disease Prediction is a Machine Learning Technique where Health care professionals used to identify the early stage of Chronic Kidney Disease and take care of their patients. We used the Hybrid Ensemble Technique to integrate multiple models derived from various algorithms. From this proposed Prediction system, we achieved about 99% accuracy. We used the Flask Web Framework to interact with the system online, where doctors and patients check their health.

5 Modules:

5.1 SQL Alchemy:

Database-In this system, we used sqlite3 database to save the users information and their login details. SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. We interact with this database using SQL Alchemy. SQL Alchemy is an open-source SQL toolkit and object-relational mapper for the Python programming language.

5.2 Flask Forms:

Forms-Forms are used to collect user inputs or to get user requests. In Flask, each form is defined as a class because of easier interaction in the server. There are four methods available to communicate between server and client. In this system, POST method is used to communicate securely.

5.3 Routes-Routing:

Routes-Routing is used to map every URLs defined in the Flask System. In other words, if the users enter the login page, the flask maps the login page to the app route to login operation defined in the system to authenticate the user. The flask routing can handle the user data dynamically sent to the server, regardless of the type of data.

5.4 User Interaction:

User Interaction-User can interact with the system by asking the admin to register their account. And, by using the login details the user can check their health. In the home page, users are able to give their data and make the system to check about the health status. The system provides its analysis in graphs also.

5.5 Prediction:

For Predicting the early stage of Chronic Kidney Disease, We used machine learning algorithms to make the prediction using different techniques. In this system, Random Forest, KNN Classifier, Logistic Regression techniques are used. Finally, Hybrid Ensemble Technique is used to integrate all ML algorithms to provide more accurate results.

In our proposed work, dataset is loaded and preprocessed to avoid noisy data from the dataset. Then feature extraction is implemented which extracts required features from dataset and avoid unwanted fields from it. Then hybrid ensemble classifier is implemented to train our training dataset. Initially KNN is used to classify our dataset once classified dataset is undergone for Random Forest (RF) then finally Logistic Regression (LR) is deployed to classify our dataset in accurate way.

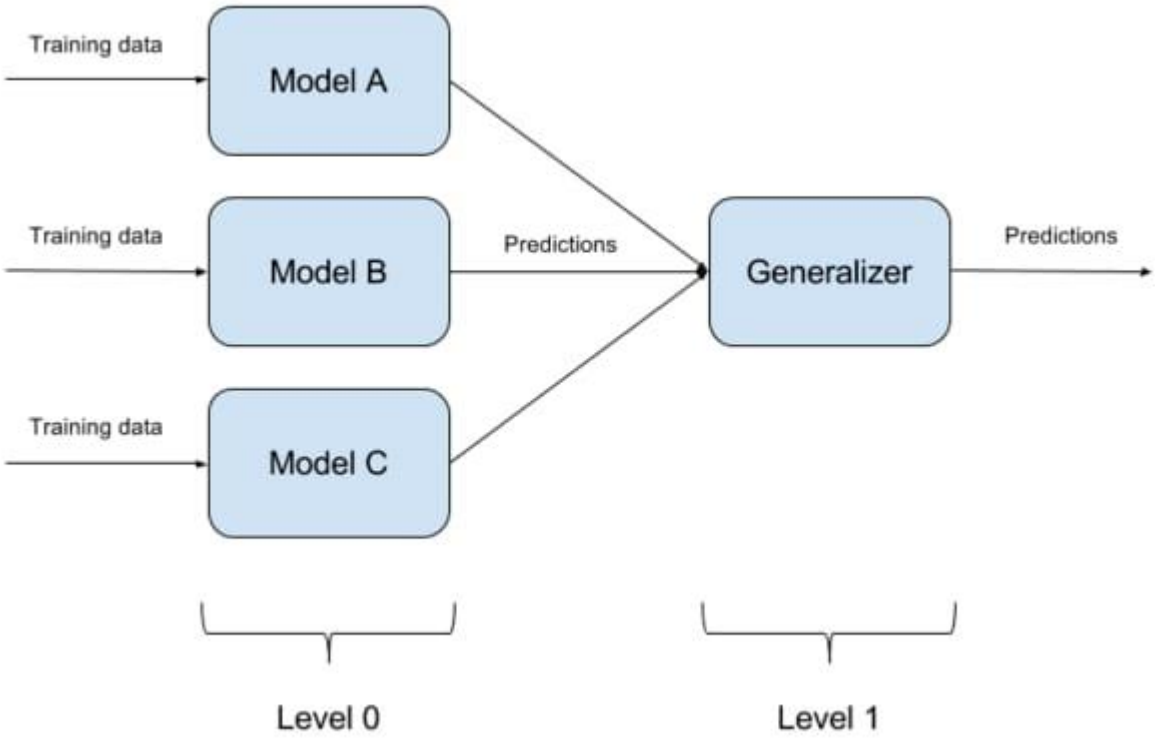

Figure 2: Working of proposed system

6 Proposed machine learning classifier

6.1 KNN classifier:

K-Nearest Neighbor (K-NN) K-Nearest Neighbor classifier is the simplest classifier method.By making use of the previously identified data points(nearest neighbor) and classified data point the classifier will detect the unidentified data point. They make use of more than one nearest neighbor for classification of data points. K-NN can be used to create early warning system in case of chronic disease. In this case the K-NN will detect the association existing between cardiovascular disease, hypertension and risk factor of different chronic diseases.This classifier can also be used to analyze heart disease patient. The data is obtained from UCI. From the experiment involving both without voting and with voting K-NN classifier the comparison result was that K-NN has accuracy without voting in diagnosis of heart diseases.

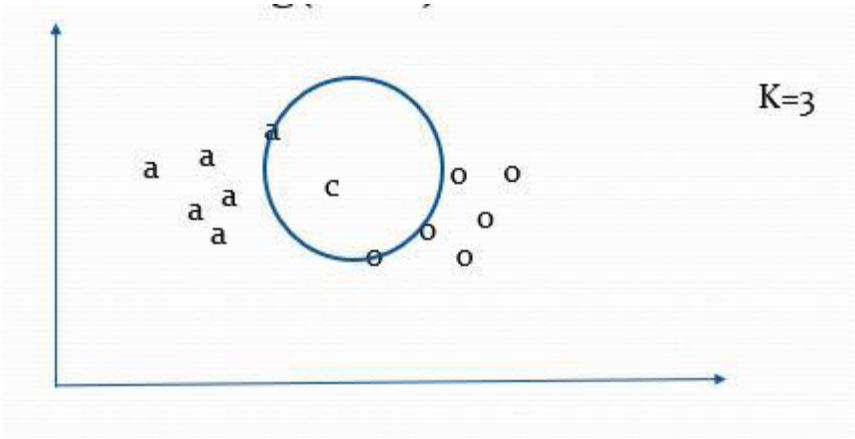

Figure 3: K-NN Classifier for Chronic Disease

6.2 Random Forest:

The random forest is an ensemble strategy that can alternatively be considered as a form of the closest predictor. Ensembles are a method used to divide and conquer performance. A collection of weak students can join forces to become a 'strong learner' as the basic premise underpinning ensemble methods. The Random Forest begins with a normal "decision tree" machine learning technique that correlates, in terms of the ensemble, to the weaker student. The algorithm of decision tree continually divides the data set into a criterion which optimizes data separation, which creates a tree-like structure. An input is entered at the top of the algorithm and the data are incorporated into smaller and smaller sets when the tree is passing through. In mixing trees with an ensemble, the random forest takes this concept to the next level. So the trees are, on the whole, weak learners and a strong learning person from the random forest. The advantages of a randsome forest classification are that it can handle unbalanced and missing data relatively quickly. The weaknesses of this approach are that it is not able to forecast outside of the extent of training data when used for regression and can overfit extremely noisy data sets.

6.3 Logistic Regression:

Logistic regression is a widely used model for analyzing the association between many, separate variables and one categorically dependent variable and the equation of the form:

$$log\,[\,p\,/\,1-p\,]\;=\;a\;+\;\beta 1\,x1\;+\;\beta 2\,x2\;+\;\cdots\;+\;\beta ixi \qquad (1)$$

Where p is the probability of an outcome of interest, β1 is an interception, ßi are β coefficients connected with each variable x , and x1..

7 Result And Discussion

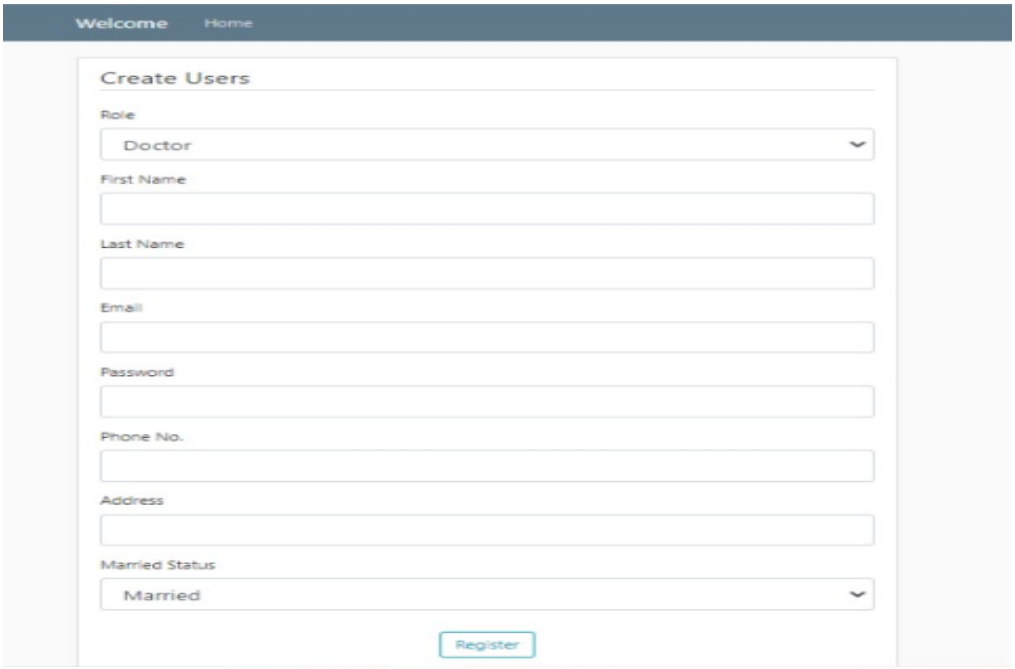

Figure 4: home page

In the above diagram it clearly shows user or doctor should enter their details for registration. This registration is used for authentication process. Hence unauthorised user access could be reduced.

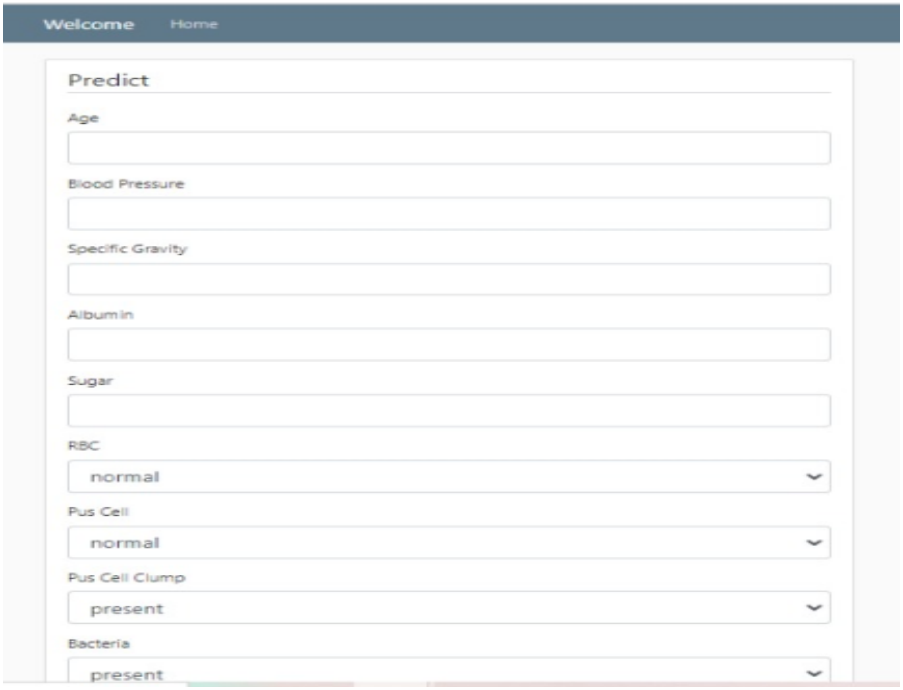

Figure 5: entering parameter for prediction

The above graph shows entering parameters like particular patient age, BP, sugar level, RBC, pulse cell and bacteria etc. Based on the inputted parameter prediction is done through hybrid ensemble technique such as KNN, RF and LR.

Figure 6: prediction of CKD

Once required parameter is entered by clicking predict option it will predict whether particular patient is suffering from CKD or not. The above figure shows that particular patient is not CKD affected patient.

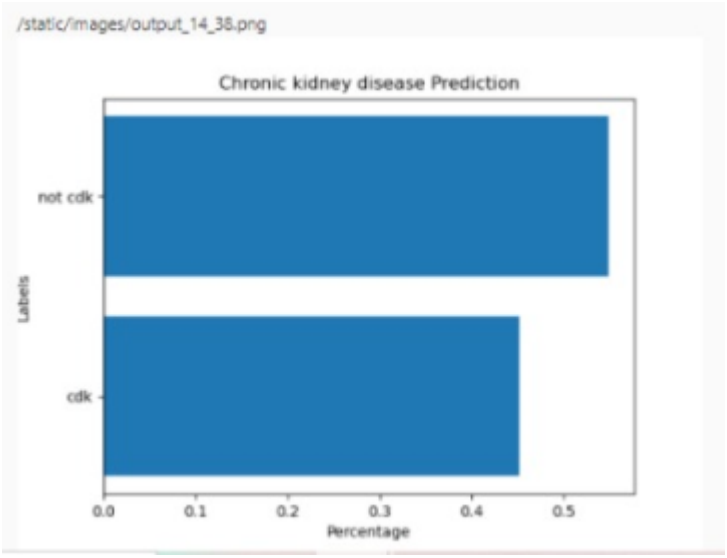

Figure 7: based on entered dataset percentage of CKD or not CKD

Based on loaded dataset it clearly shows percentage of CKD and not CKD hence the prediction accuracy is high because of hybrid approach implementation.

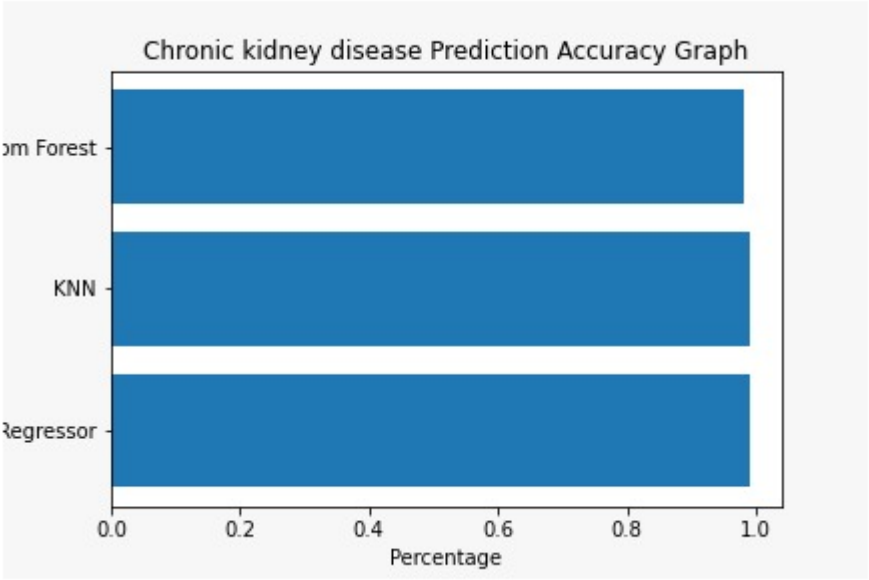

Figure 8: CKD prediction accuracy graph

Three different machine learning algorithm is implemented together as hybrid approach in prediction of CKD. Hence the accuracy level of prediction is shown clearly in above figure.

8 Conclusion

Chronic kidney disease identification is a difficult task that should be done in accurate way. More number of research works is going on in detection CKD and currently Machine learning algorithms are also used. However machine learning algorithms achieve accuracy in detection but not up to the mark of accuracy. Hence in our proposed work, hybrid ensemble technique was implemented which is a combination of KNN, RF and LR which attains maximum accuracy. The trained output result of KNN, RF and LR are combined together is represented as Generalization hence outcome of these combined result is high in accuracy. Hence our results

shows that our proposed method achieves maximum accuracy in detection of CKD compared to other machine learning algorithms.

References

[1]. Bidri Deepika*, Vasudeva Rao KR, Dharmaj N Rampure, Prajwal P and Devanand Gowda G. "Early Prediction of Chronic Kidney Disease by using Machine Learning Techniques" American Journal of Computer Science and Engineering Survey.

[2]. Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A, "Chronic Kidney Disease Prediction using Machine Learning" International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 07 (July 2020).

[3]. Sai Prasad Potharaju and M.Sreedevi, "Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data" Journal of Engineering Science and Technology Review 9 (5) (2016) 201- 207.

[4]. Shubham Gupta, Vishal Bharti, Anil Kumar, "A Survey on various Machine Learning Algorithms for Disease Prediction" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6C, April 2019.

[5]. Sahil Sharma, "A Two Stage Hybrid Ensemble Classifier Based Diagnostic Tool For Chronic Kidney Disease Diagnosis Using Optimally Selected Reduced Feature Set" International Journal of Intelligent Systems and Applications in Engineering 2018.

[6]. Sirage Zeynu, Shruti Patil, "Prediction of Chronic Kidney Disease Using Feature Selection and Ensemble Method" International Journal of Pure and Applied Mathematics Volume 118 No. 24 2018.

[7]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[8]. Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE International Conference on Computer Software and Applications (COMPSAC), IEEE, Jul. 2017, doi: 10.1109/COMPSAC.2017.84

[9]. Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," July18, 2016.

[10]. 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.