IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

Classification of Chronic Kidney Disease using Combination Feature Selection Techniques and Classifiers

A. K. Shrivas[1*], Sanat Kumar Sahu[2]

[1]Dept. Of IT., Dr. C. V. Raman University, Bilaspur (C.G.), India
[2]Dept. Of Computer Science, Govt. Kaktiya P.G. College, Jagdalpur (C.G.), India

[*]*Corresponding Author: sanat.kosa1@gmail.com, Tel.: +919770873106*

Available online at: www.ijcseonline.org

***Abstract*—** The aim of the study is to predict significant features from dataset of Chronic Kidney Disease features. It represents the data in a tabular and graphical manner to form its clear understanding. This investigation helps in the crucial role of features and experimental features in the CKD dataset and their associations, their dependability for coming up with any classification system. It also shows that how CKD can be diagnosis by exploiting data mining techniques. The Data Mining algorithm is an inspirational force in detecting abnormalities in various data sets and with a good success utilized in various classification and feature selection task. The different kinds of Decision Tree-based classifiers like RF (Random Forest), J48 (C4.5), C5.0, and CART (Classification and Regression Tree) and their ensemble model have experimentally validated CKD dataset and our result is evaluated. Our result representation that the ensemble models classifier reached the most favourable performances on the identification of CKD dataset before and after the feature selection.

***Keywords*—** Chronic Kidney Disease, Feature Selection Techniques, Classification, Random Forest, J48(C4.5), C5.0, CART, Ensemble model, Genetic Search, Greedy Stepwise

I. INTRODUCTION

Nowadays Chronic Kidney Disease (CKD) is increasing, mostly in all high-income and middle-income, and also in several low-income countries [1]. The CKD may be a burden to peoples, families, and society. The CKD issues alter the function and structure of the kidney irreversibly, over months or years [2]. Early detection of CKD patients is crucial when treatment can potentially reverse, delay, or prevent progression of the disease. The different data mining [3] classification approach and machine learning algorithm applied for prediction of chronic kidney diseases. Classification may be a method that's usually utilized in data processing and is used to find hidden patterns within the database. Classification is employed to insert data objects into many predefined classes. Well-defined characteristics play an important role in the performance of classification. The data classification relies on a learning algorithm. Training cannot be done by exploitation of all the data. This is often done on the data sample concerning to the data collection. The aim of learning is to build a classification model [3]–[5]. Feature selection techniques [6], on the other hand, choose the foremost informative features from the original dataset. So there is no damage to their physical explanation. The fundamental subject of the feature selection is to find the most relevant features from thousands of related ones in a specific area [7]. It conjointly helps in increasing execution speed and accuracy of classification algorithms.

II. RELATED WORK

Various researchers have studied and investigated within the field of health care that involve numerous Chronic Diseases like Cancer, Heart, and other diseases diagnosis. Many authors have studied concerning Chronic Kidney Diseases (CKD). The [8]) used four classification algorithms like Random Forest (RF), Classification and Regression Tree (CART) and Support Vector Machine (SVM) to classify the CKD and propose an ensemble model. The proposed ensemble model with proposed UBFST offered superior accuracy compared to others existing FSTs and all individual classifiers. Polat, Danaei Mehr, & Cetin (2017) included two kinds of feature selection approach, i.e., wrapper and filter approach are adapted to diagnose CKD. The experimental results evidenced that Support Vector Machine (SVM) classifier has used filtered subset evaluator with the BFS engine feature selection style offered an enhanced accuracy rate (98.5%) in the classification of CKD. [9] covered classifiers like Artificial Neural Network, Support Vector Machine, k-Nearest Neighbor, C4.5 and Random Forest in favour of identification of CKD. The Random forest (RF) classifier got maximum performance on the identification of CKD. [10] have predicted CKD problems using completely

dissimilar machine learning algorithms like Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decision Tree (C4.5), Bayesian Network (BN) and K-Nearest Neighbour (K-NN). The investigational results revealed that the MLP and C4.5 lead and reasonable, the ROC curve, the C4.5 has the best result. [11] used clinical data for the identification of CKD. They included machine learning algorithms like K-nearest neighbours (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers. The results explained that the SVM classifier provides the uppermost accuracy, sensitivity later than training and testing by the proposed technique. [12] used three completely different classification techniques similar to Back Propagation Neural Network, Radial Basis Function and Random Forest for classification of CKD. Radial basis function network provides the uppermost accuracy of 85.3%. [13] have investigated three classification techniques i.e. Naïve Bayes, J48 and SMO and demonstrate of accuracy. J48 classifier achieved the most effective classification accuracy evaluated up to others. [14] have used different classifiers like Random Forest (RF), Sequential Minimal Optimization (SMO), Naive Bayes, Radial Basis Function (RBF) and Multilayer Perceptron (MLP), Simple Logistic (SLG) techniques for the predictions task of CKD. The Random forest achieved higher performance compared to other classifiers.

III. METHODOLOGY

In this section, we have used four classification techniques based on decision tree and their ensemble models for classification of CKD dataset.

- **Random Forest (RF):** Random Forest (RF) [15], [16] is a together classifier that can be found in the many decision trees [3] and outputs the class with the purpose of is the mode of the classes output through individual trees. Random Forests are frequently used at the time we have especially vast training datasets and an awful number of input variables (hundreds or even thousands of input variables). A random forest model consists of tens or hundreds of decision-making trees.

- **Classification and Regression Tree (CART):** CART [3], [5] may be a non-restrictive DT learning technique to facilitate assembles whichever classification or regression trees, betting on whether or not the dependent variable is categorical or numeric. It constructs a binary DT by uninflected the record at each node, in step with a gathering of a single attribute. CART bring into play the Gini index to establishing the most effective divide applied statistical techniques soak up typically utilized in health care in support of the classification of varied diseases.

- **J48 (C4.5):** C4.5 [3], [17] is an algorithm used to produce a decision tree developed by Ross Quinlan. C4.5 is implementing a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. The decision trees [5] produced by C4.5 be capable of use for classification, and for this rationale, C4.5 is often referred to as a statistical classifier.

- **C5.0:** This is a decision tree supported classifier developed by Ross Quinlan and is an extension of C4.5. It while not human intervention extracts classification rules in the form of the decision tree from specified training data. C5.0 [3], [17] has several benefits over C4.5 in provisions of time and memory space required; the tree generated by C5.0 additionally terribly tiny as compared to the C4.5 formula that ultimately improves the classification accuracy.

- **Ensemble Models:** An ensemble model or hybrid model [3], [18], [19] is an amalgamation of two or more skilled individual's classifiers and creates a new composite model. It can be used to reduce the error of any weak learning algorithm. The main purpose of combining these individual classifications is to develop an ideal model that will improve the performance of each individual's classifier.

- **Feature Selection Techniques (FST):** Feature Selection technique is also called Attribute selection techniques or Feature Reduction techniques. The feature selections wherever identify the evaluation method and search method.

Evaluation Methods: Subset Evaluation and Single Attribute Evaluation are the two frequently used techniques in feature estimation area [4].

CfsSubsetEval*:* Consider the predictive value of each attribute individually, along with the degree of redundancy among them [4].

ClassifierSubsetEvaluator*:* Uses the classifier specified in the object editor to estimate the set of features on the training data or on the separate set of holdouts[4].

Search Methods: Search algorithms [4], [20] are a unit essential for feature selection for the reason that it affords some way to search for attributes.

- **Genetic Search:** Performs a search using the simple genetic algorithm described by Goldberg [21], [22]. The parameter includes populace size, number of generations, and probabilities of crossover and mutation [4]. One'll be able to specify a listing of attribute index because the start line, that becomes a member of the initial

population. Progress reports are often generated each generation most.

- **Greedy Stepwise Search:** When beginning with an empty set, it selects the variable by further selection and eliminates the useless variable by selecting backwards to find the most effective feature subset. Throughout the search method, a replacement assortment of candidate feature sets was created by adding different features to the most effective feature subset. When evaluating all the subsets, the most effective feature set was elect. The algorithm continues until the fresh generated archive of the set doesn't exceed the most effective current subset [4].

DATA SET: These follow a line of investigation of attention on the classification of Chronic Kidney Disease (CKD). The CKD [23] dataset is collected from UCI repository having 24 features, 400 instances and 1 class binary nature (ckd or notckd). The features of data have numerical and nominal value.

IV. RESULTS AND DISCUSSION

The investigational work is done in two sections. In the first section, classification of CKD is analyzed and developed, whereas in the second section, there is a reduction in the features of the CKD data set to get better performance of models. The results of 4 decision tree techniques and their ensemble models are evaluated within the experiments with all features of CKD dataset. All decision tree techniques are evaluated by the proposed technique. During this experiment, the 10-fold cross validation is employed to classify all models and therefore the average results are shown in figure 1.

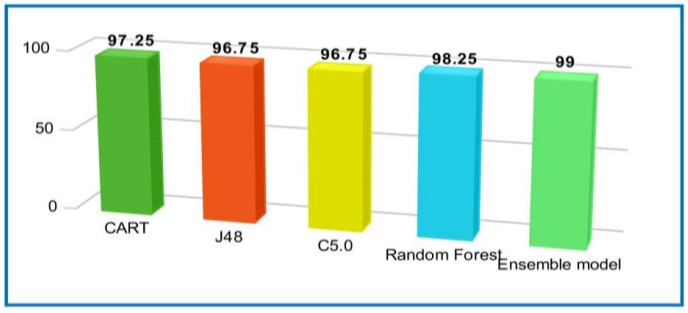

Figure 1:. Accuracy of different classifications models

In the figure one showed the accuracy of the various classification models within the type bar charts. It was evident from the actual fact that the J48 and C4.5 have 96.75%, CART has 96.75% and also the Random Forest has 98.25 %accuracy. The accuracy of the ensemble model is 99.00% that was offer the highest accuracy compared to other classifier.

Table 1. Features of CKD dataset Selected by different Search Method with CfsSubsetEval evaluator method

Search Method	Total Selected Features	Selected of Features
Genetic Search	14	2,3,4,6,10,12,13,15,16,18,19,20,22,23
Greedy Stepwise	14	2,3,4,6,10,12,13,15,16,18,19,20,22,23

Table 2. Features of CKD dataset Selected by different Search Method with ClassifierSubsetEval (Classifier J48) evaluator method

Search Method	Total Selected Features	Selected of Features
Greedy Stepwise	5	3,4,6,12,15
Genetic Search	7	3,4,6,8,16,17,20

The ensemble models are evaluated within the experiments once the feature selection techniques have applied. During this experiment, the 10-fold cross validation is employed to classify the ensemble models and therefore the average results are shown in table 3.

Table 3.Accuracy of ensemble models

FST applied on the best model (Ensemble model RF, J48, C5.0, CART)			
Sr. No.	Search method	Accuracy	Total Number of Features
1	Greedy Stepwise	99.75	14
2	Genetic Search	99.75	14
3	Greedy Stepwise	99.75	05
4	Genetic Search	99.00	07

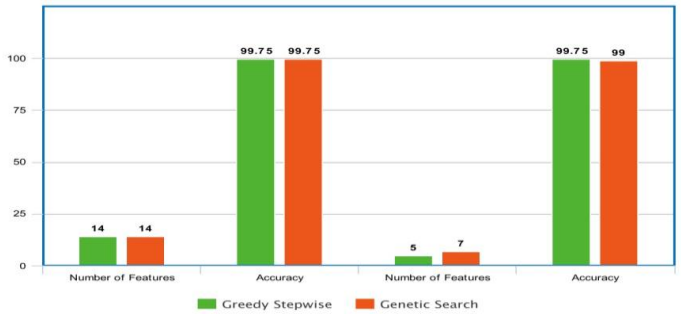

Figure 2: Accuracy of ensemble models with different feature Selection Techniques

The figure 2 shows the accuracy of the ensemble model and selected features bar chart. The lowest accuracy of the ensemble model is 99.00% which comes with has 07 features. It is evident from the accuracy of the ensemble model have 99.75% which comes with have 14 features. Also the accuracy of ensemble model is achieved 99.75% with 05 features. So, the best accuracy of our ensemble model is 99.75%, it has least 05 features. It has least features with the best accuracy.

V. CONCLUSION and Future Scope

In the above model, Decision tree Random Forest, CART, J48 (C4.5) and C5.0 and their ensemble model classifiers used to predict chronic kidney disease. The CKD datasets are classified employing a combination of feature selection and classifier algorithm. Observed results achieved by conducting experiments on CKD data sets granting for four decision tree based classifier methods; one proposed ensemble model two feature selection algorithms supported the accuracy of this criterion. From the experimental results, it is often seen that the ensemble model classifier offers the very best accuracy in all cases. The ensemble model is reasonably efficient and best suitable for identification of CKD issues.

REFERENCES

[1] M. P. Webster A.C., Nagler E.V., R.L., "Chronic Kideny disease," 2016. [Online]. Available: http://www.worldkidneyday.org/faqs/chronic-kidney-disease/. [Accessed: 28-Feb-2018].

[2] M. S. McManus and S. Wynter-Minott, "Guidelines for Chronic Kidney Disease: Defining, Staging, and Managing in Primary Care," J. Nurse Pract., vol. 13, no. 6, pp. 400–410, 2017.

[3] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third. Elsevier, 2012.

[4] I. H. Witten, E. Frank, and M. A. Hall, Data mining. 2011.

[5] A. Pujari, Data mining techniques, Third. University press, 2013.

[6] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, no. 3, pp. 1157–1182, 2003.

[7] A. Bhalla and R. K. Agrawal, "Microarray gene-expression data classification using less gene expressions by combining feature selection methods and classifiers," Int. J. Inf. Eng. Electron. Bus., vol. 5, no. 5, pp. 42–48, 2013.

[8] A. K. Shrivas, S. K. Sahu, and H. S. Hota, "Classification of Chronic Kidney Disease with proposed Union Based Feature Selection Technique," no. 2007, pp. 503–507, 2018.

[9] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of Chronic Kidney Disease by Using Random Forest," C. 2017 Proc. Int. Conf. Med. Biol. Eng. 2017, vol. 7, no. 1, pp. 589–594, 2017.

[10] B. Boukenze, A. Haqiq, and H. Mousannif, "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques," vol. 397, 2017.

[11] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Manag. Innov. Technol. Int. Conf., p. MIT-80-MIT-83, 2016.

[12] D. N. R. S.Ramya, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 1, pp. 812–820, 2016.

[13] M. Arora and E. A. Sharma, "Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka," Int. J. Comput. Appl., vol. 6, no. 4, pp. 20–26, 2016.

[14] M. Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 2, pp. 24–33, 2016.

[15] R. Nagalla, P. Pothuganti, and D. S. Pawar, "Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests," Procedia Comput. Sci., vol. 109, no. 2016, pp. 474–481, 2017.

[16] R. Parimala and R. Nallaswamy, "A Study of Spam E-mail classification using Feature Selection package," Glob. J. Comput. Sci. Technol., vol. 11, no. 7, pp. 45–54, 2011.

[17] A. K. Shrivas, S. K. Sahu, and S. K. Singhai, "Decision support system for classification of chronic kidney disease with principle component analysis," vol. 14, no. 2, pp. 105–110, 2017.

[18] Sivanandam and Deepa, Principles of Soft Computing, Second. wiley, 2014.

[19] S. Haykin, Neural Networks and Learning Machines, vol. 3. 2008.

[20] C. Arun Kumar, M. P. Sooraj, and S. Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets," Procedia Comput. Sci., vol. 115, pp. 209–217, 2017.

[21] D. E. Goldberg, "Genetic Algorithms in Search Optimization & Machine Learning." p. 412, 1989.

[22] M. Hall, "Correlation-based Feature Selection for Machine Learning," Methodology, vol. 21i195-i20, no. April, pp. 1–5, 1999.

[23] "UCI Machine Learning Repository of machine learning databases," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed: 01-Jan-2016].

Authors Profile

Dr. Akhilesh Kumar Shrivas is working as Assistant Professor in Department of Information Technology, Dr. C.V. Raman University, Bilaspur, India. He obtained his Master's Degree in Computer Application from Guru Ghasidas Vishwavidyalaya, Bilaspur, India and Ph. D. in Computer Science from Dr. C.V. Raman University, Bilaspur, India. He has 6 year research experience and published more than 50 research papers in reputed journals and conference proceedings and attended workshop and conference at national and international level. His research interests are data mining, soft computing, big data and information security.

Mr. Sanat Kumar Sahu is working as Assistant Professor in Department of Computer Science, Govt. Kaktiya PG College, Jagdalpur (Bastar) Chhattisgarh, India. He obtained his Master's Degree in Computer Application from Guru Ghasidas Vishwavidyalaya, Bilaspur, India and M. Phil in Computer Science from Dr. C.V. Raman University, Bilaspur, India. He has more than 7 years teaching and 02 years research experience. He has published more than 12 research papers in reputed journals and attended workshop and conference at national and international level. His area of interest includes soft computing, machine learning, and data mining.

