A Novel Ensemble Model for Classification of Chronic Kidney Disease With Selected Features and Components

Sanat Kumar Sahu

Department of Computer Science, Govt. Kaktiya P.G. College Jagdalpur (C.G.), India

Author's Mail Id:sanat.kosa1@gmail.com

Abstract— Diagnosis of health conditions is an incredibly difficult and significant issue in the field of medical science. Classification, dimension reduction technique (DRT), feature selection techniques (FST) play a very important role in the quick and accurate identification of disease. The chronic kidneys disease (CKD) dataset is connected into three classification methods like RF, J48 and C5.0. The proposed ensemble model (RF, J48 and C5.0) gives better accuracy i.e. 99.75% contrast with all classifiers with selected feature subset. All classification models give a better outcome with proposed PC-DRT and GA-FST when contrasted with without FST. The outcomes showed that utilizing GA-FST has computationally enhanced the classification accuracy.

Keywords— Classification, chronic kidneys disease, dimension reduction technique, ensemble model, feature selection techniques, genetic algorithm, principal component analysis

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a universal public health problem, affecting around 10% of people worldwide [1]. Yet, there is very little evidence about however CKD is often diagnosed efficiently and automatically. The CKD or Chronic Renal Disease (CRD) slowly progresses and is hard to identify and the kidney loses its functionality. Normally it may not be detected before a 25% loss of kidney functionality. The beginning of kidney failure of patients cannot be recognized because kidney failure may not give any symptoms initially [2]. Nowadays healthcare and medical sciences have involved several innovative computer technologies for useful and efficient diagnostic techniques. The clinical treatments and diagnosis system are time-sensitive, without the compromising of least time constraints one has to do medical diagnosis properly. The classification techniques are acceptable to be an appropriate methodology with the purpose of effective diagnosis in medical science and healthcare. FST is a crucial technique to explain the problem of dimensionality in machine learning by choosing appropriate and non-redundant features[3], [4]. The study aims to propose a model for the identification of CKD using classifications, FST and DRT techniques. The CKD data sets collected from the UCI machine learning repository have been used for the present study.

II. RELATED WORK

This research begins with concise related work in the field of CKD.
Tsai et al.(2018)[5] proposed a data clustering algorithms framework for implementing k-means, genetic k-means and particle swarm optimization on a standalone system and spark. The result of the proposed framework makes the cloud computing surroundings treasured for developing clustering algorithms. Thomas et al. (2018)[6] worked on clustering algorithms and attribute extraction methods to reveal beneficial tendencies in industrial chemical process data. The comparison of cluster results against actual labels in process knowledge and supervised clustering matrix data is executed to compare the overall performance of one-of-a-kind combinations of dimensions and data clustering approaches. Najafabadi et al. (2017) [7] compared different prominent deviations of the preliminary CF and CF techniques, which incorporated potential education through K-Means clustering method, Hybrid representation method with public datasets, i.e. million hear datasets. The result confirmed that the proposed approach achieved the upper performance in contrast to the essential CF and different extended versions of CF methods in terms of Precision, Recall metrics, even once the data is incredibly sparse. Hou & Liu (2017)[8] had proposed the constraint indented clustering method supported on the dominated set methods and clustering integration. The experimental result confirmed that it is more valuable than already existing clustering techniques. Subasi et al., (2017) [9] used Machine learning methods like Artificial Neural Network (ANN), Support Vector Machine (SVM), K-nearest neighbors (K-NN), C4.5, and Random Forest (RF) and validated with CKD dataset. The outcome showed that the RF classifier performed outstandingly. Kunwar et al. (2016) [10] used two types of classification techniques like Naive Bayes and Artificial Neural Network (ANN) to classify Chronic Kidney Disease (CKD). They experimented with the RapidMinor tool. The outcomes show that Naive Bayes produce more

accurate results than ANN. Sinha (2015) [11] had used two classification methods namely Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) for the classification task of the CKD dataset. The outcome was clear that the KNN performed better than SVM.

III. METHODOLOGY

This research begins with concise related work, then a description of benchmark dimension reduction technique (DRT), a description of our methodology in the research work, and the results obtained by using the CKD datasets.

A. Random Forest: Random forests (RF) or random decisions are a hybrid learning technique for classification and regression and other functions, which, at training time, is that the technique of developing a massive range of trees and category/class (classification) or mean prediction (regression) of trees [12].

B. C4.5 (J48) and C5.0 Trees
C4.5 [13] could be a methodology utilizes to provide a decision tree developed by Ross Quinlan.C4.5 is a related expansion of Quinlan's former ID3 algorithm. The decision trees created by C4.5 can be conveyed into contributing for classification, and for this rationale, C4.5 is frequently referred to as a statistical classifier. C 5.0 is an associate extension of Quinlan's earlier C4.5 algorithm. C4.5 prepared a come-to of enhancement to ID3.

Figure 1: C 4.5 decision tree

C. Principal Component Analysis (PCA): The Principal Component Analysis (PCA)[14] is an approach to decrease the dimensionality of a dataset set consisting of a huge number of interrelated variables while keeping as a lot as feasible of the version existing in the dataset[15], [16]. This is carried out by remodeling to a brand new set of variables, the principal components (PCs), that are unrelated, and that are ordered so the primary ordered so that the first few maintain most of the variant current in all of the unique variables or PCA seriously change in new attribute space, have been aspects are uncorrelated and ranked by way of degree are explaining through variance [17][18].

D. Genetic Algorithm: Genetic Algorithms (GA) [19], [20] is discovered to mimic numerous of the method observed in natural evolution. The basic concept of GA is used to the command of evolution to solve optimization problems. The earliest introduction of the unique Genetic Algorithm was by John Holland who pretends it in the early 1970s. Genetic algorithms [21] are adaptive search techniques based on the principles of natural selection in biology. They use a population of competing solutions—changed over time—to converge to an optimal solution [22].

E. DATASET
In this research, the use of CKD data sets collected from the UCI machine learning repository [23] has been used.

IV. RESULTS AND DISCUSSION

In the initial part, we classified the CKD dataset into three classification methods like RF, C5.0, and J48. Classification with 10-fold cross-validation technique and verify the accuracy of all classifiers. Then proposed an ensemble model, the ensemble model has three classifiers namely RF, C5.0 and J48 which are combined through the stacking ensemble technique. After the FST and DRT are applied the selected features ensemble model and got the accuracy.

Table 1: Classifier and their accuracy

S. No.	Name of algorithms	WFST
1	RF	98.25
2	C5.0	96.75
3	J48	96.75
4	RF+J48+C5.0	98.00

Table 1 shows the experimental results of RF, C5.0, J48 and their ensemble model. The result shows that the proposed ensemble model gives better performance compared to the individual model.

Figure 2: Comparison of classifiers accuracy

Figure 2 shows the experimental results of RF, C5.0, J48 and their ensemble model (RF, C5.0, J48). The column char shows that the proposed ensemble model gives better accuracy but the highest accuracy gives the individual model RF.

Table 2: Ensemble classifier and their accuracy

S. No.	Name of algorithms	WFST	PC-DRT	GA-FST
1	RF+J48+C5.0	98.00	99.50	99.75

Table 2 also compared the performance of the proposed ensemble model classifier in terms of accuracy and the number of features and components used in the CKD dataset. After applied the FST and DRT it is clear that the accuracy of the proposed ensemble model is enhanced compared to all features of CKD.
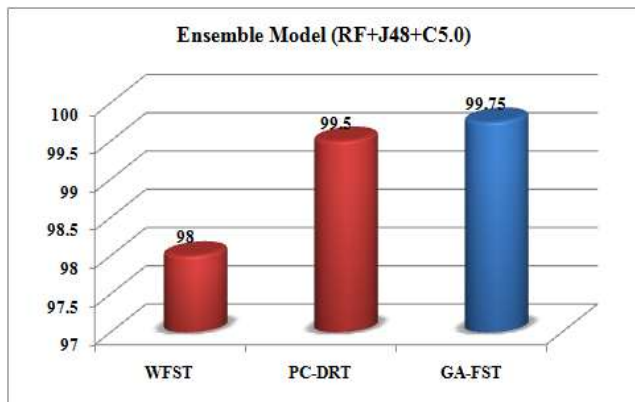
Figure 3: Comparison of classifier accuracy

Figure 3 shows that the proposed ensemble model gives higher accuracy of 1.50% compared to WFST and PC-DRT. The same way ensemble model gives higher accuracy of 1.75% compared to WFST and GA-FST.

V.   CONCLUSION AND FUTURE SCOPE

Diagnosis of health conditions is an incredibly difficult and significant issue in the field of medical science. Classification, Dimension reduction technique (DRT) feature selection techniques (FST) play a very important role in the quick and accurate identification of disease. The CKD dataset is connected into three classification methods like RF, J48, and C5.0. The individual model and ensemble model (RF, C5.0, J48) gives better accuracy compared to all features and reduced features of CKD. The best classification model is obtained through experimental work where the proposed ensemble model (RF, C5.0, J48) is applied to FST and DRT. The ensemble model (RF, J48, and C5.0) gives better accuracy i.e. 99.75% contrast with all classifiers with selected feature subset. All classification models give a better outcome with proposed PC-DRT and GA-FST when contrasted with without FST. The outcomes showed that utilizing GA-FST has computationally enhanced the classification accuracy.

The best classification model obtained through experimental work is further used to apply the feature selection technique (FST) to remove irrelevant features from the CKD dataset.
.

REFERENCES

[1] 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.**

[2] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," J. Med. Syst., **vol. 41, no. 4, 2017.**

[3] N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," Complex Intell. Syst., no. **March, 2018.**

[4] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., **vol. 1, no. 3, pp. 131–156, 1997.**

[5] C. W. Tsai, S. J. Liu, and Y. C. Wang, "A parallel metaheuristic data clustering framework for cloud," J. Parallel Distrib. Comput., **vol. 116, pp. 39–49, 2018.**

[6] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," J. Process Control, **vol. 67, pp. 160–175, 2018.**

[7] M. K. Najafabadi, M. N. ri Mahrin, S. Chuprat, and H. M. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," Comput. Human Behav., **vol. 67, pp. 113–128, 2017.**

[8] J. Hou and W. Liu, "Parameter independent clustering based on dominant sets and cluster merging," Inf. Sci. (Ny)., **vol. 405, pp. 1–17, 2017.**

[9] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of Chronic Kidney Disease by Using Random Forest," **vol. 7, no. 1, pp. 589–594, 2017.**

[10] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease Analysis Using Data Mining Classification," Cloud Syst. Big Data Eng. (Confluence), 2016 6th Int. Conf. IEEE, **pp. 300–305, 2016.**

[11] P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," **vol. 4, no. 12, pp. 608–612, 2015.**

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

[13] J. Ross, Q. Morgan, and K. Publishers, "Book Review : C4 . 5 : Programs for Machine Learning," **vol. 240, pp. 235–240, 1994.**

[14] Y. Zhang and L. Wu, "Classification of fruits using computer vision and a multiclass support vector machine," Sensors (Switzerland), **vol. 12, no. 9, pp. 12489–12505, 2012.**

[15] L. C. Paul, A. Al Suman, and N. Sultan, "Methodological Analysis of Principal Component Analysis (PCA) Method," IJCEM Int. J. Comput. Eng. Manag. ISSN, **vol. 16, no. 2, pp. 2230–7893, 2013.**

[16] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease," Int. J. Mach. Learn. Comput., **vol. 5, no. 4, pp. 258–263, 2015.**

[17] M. Einasto et al., "SDSS DR7 superclusters. Principal component analysis," **pp. 1–13, 2011.**

[18] P. Verma, V. K. Awasthi, and S. K. Sahu, "A Novel Design of Classification of Coronary Artery Disease Using Deep Learning and Data Mining Algorithms," Rev. d ' Intell. Artif., **vol. 35, no. 3, pp. 209–215, 2021.**

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

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

[21] P. J. Fleming and C. M. Fonseca, "Genetic Algorithms in Control Systems Engineering," IFAC Proc. **Vol., vol. 26, no. 2, pp. 605–612, 1993.**

[22] S. K. Sahu and A. K. Shrivas, "Analysis and Comparison of Clustering Techniques for Chronic Kidney Disease With Genetic Algorithm," Int. J. Comput. Vis. Image Process., **vol. 8, no. 4, pp. 16–25, 2018.**

[23] "CKD," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed: 01-Jan-2016].

AUTHORS PROFILE

Dr. Sanat Kumar Sahu is working as an Assistant Professor in the Department of Computer Science, Govt. Kaktiya PG College, Jagdalpur (Bastar) Chhattisgarh, India. He has more than 11 years teaching Experience. His area of interest includes soft computing, machine learning, and data mining.. He has more than 22 research paper in national and international journals.