Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290

Chronic Kidney Disease (CKD) Prediction Using Supervised Data Mining Techniques

S. Rajarajeswari
PG and Research Department of Computer Science, M.G.R College, Hosur,
Krishnagiri-635002, Tamilnadu.
Email: rajeswariannam@gmail.com
T. Tamilarasi
PG and Research Department of Computer Science, M.G.R College, Hosur,
Krishnagiri-635002, Tamilnadu.
Email: thavasitamil@gmail.com

----------**ABSTRACT**----------

Diseases are causing high rates of mortality in the modern world, chronic kidney disease (CKD) is one of the major causes of mortality, and it has a long-term disability. The predisposing factors for CKD include diabetes mellitus, hypertension, cardiovascular diseases, smoking, obesity, family history of kidney disease and congenital kidney problems. CKD is associated with many complications such as, proteinuria, anaemia of CKD, CKD-mineral and bone disorder, dyslipidemia and electrolytes imbalance. Renal replacement therapy (dialysis and kidney transplantation) is the treatment of choice for CKD. Data mining is an accurate technique helps to predict the disease using various methods includes logistic regression, naive bayes classification, k-nearest neighbours, and support vector machine. Apart from these previous techniques, it was necessary to use a classification method for data segmentation according to their diagnosis and regression method for finding risk factors. In this present study, data are classified using proposed Identification of Pattern Mining, Decision Tree methods and regression techniques are used to obtain the best levels and this can be taken as metrics that the proposed methods can help in diagnosing a patient with CKD.

Keywords - **Chronic Kidney Disease, CKD, Data Mining, Identification of Pattern Mining, Decision Tree.**

Date of Submission: May 23, 2021 Date of Acceptance: Jun 17, 2021

I. INTRODUCTION

In the recent era diseases and its effects are increasing day by day. The aim of data mining is to make sense of large amounts of mostly supervised data and unsupervised data, in some domain [1]. Chronic kidney disease (CKD) also known as chronic renal disease, is a gradual loss in renal function over a period of an inordinate length of time. The symptoms of worsening kidney function are unspecific, and might include feeling generally unwell and experiencing a reduced appetite (feeling Hunger). The reported prevalence of CKD in India is varying from <1% to 13%, and some studies also reported higher prevalence of 17% [2]. Diabetes and Hypertension are the most common cause of CKD, other causes include Autoimmune disorders (such as SLE and scleroderma), Birth defects of the kidneys (such as polycystic kidney disease), Certain toxic chemicals, Glomerulonephritis, Injury or trauma, Kidney stones and infection problems with the arteries leading to or inside the kidneys, Analgesics and other nephrotoxic drugs, Reflux nephropathy, and other kidney diseases. In this proposed work, classification algorithms like Identification of Pattern Mining (IPM), and decision tree (DT) methods for the diagnosis of chronic kidney diseases. The Identification of Pattern Mining outperformed over other techniques.

The rest of the paper is followed as: section 2 represents review on various disease life threatening issues, section 3 represents related works in diagnosis of chronic kidney disease, contains dataset and methods used. Section 4 represents the results and discussion. Section 5 comprises conclusion and at last references are mentioned.

II. RELATED WORK

Masethe and Masethe, used J48, REPTREE, Naïve Bayes, Bayes Net, Simple CART to predict heart disease [3]. In their study, the authors used Kappa Statistics, Mean Absolute Error, Root Mean Squared, Relative Absolute Error and Root Relative Squared Error, for the analysis. The accuracy of the prediction was 99.07%, 99.07%, 97.22%, 98.14% and 99.07% in J48, REPTREE, Naïve Bayes, Bayes Net and CART, respectively.

A study by Patil showed that neural network trained with the selected patterns for effective prediction of Heart Attack. The author used K-mean clustering algorithm on the pre-processed data [4].

P.K. Anooj et al. proposed system for the diagnosis of heart disease [5]. This study was carried out from the computerized approach for generation of weighted fuzzy rules and decision tree, creating a fuzzy rule-based decision support system from raw dataset (UCI).

Manikandan. T et al. excerpt the item set relations by using association rule. The data classification was based on MAFIA algorithms which resulted in better accuracy. The most common techniques were used to evaluate the

Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290

data as entropy based cross validation and partition techniques and the results were compared. MAFIA (Maximal Frequent Itemset Algorithm) used a dataset with 19 attributes and the goal of the research work was to have highly accurate recall metrics with higher levels of precision [6].

Hybrid Intelligent techniques for the prediction of heart disease were presented by R. Chitra et.al. Some Heart disease classification system was reviewed in this study and concluded with justification importance of data mining in heart disease diagnosis and classification. The classification accuracy can be improved by reduction in features [7].

A hybrid model for classifying Pima Indian Diabetic Database (PIDD) was developed by Jayaram et al. The author defined model consists of two stages.

- The K-means clustering was used to identify and eliminated incorrectly calc silicate instances.
- A fine-tuned classification was done using Decision tree C4.5 by taking the correctly clustered occurrence of initial stage.

Trial results signify that flowed K-means clustering and the rules generated by flowed C4.5 tree with definite data is easy to understand as compared to rules generated with C4.5 alone with continuous data. The cascaded model with categorical data obtained [8].

These studies, Decision tree and ID3 algorithms attained 72% and 80% of accuracy respectively. Han et al., proposed Data mining techniques through Rapid Miner for the classification of diabetes data analysis and diabetes prediction model [9].

The data were pre-processed and discretized the data by Breault et al. rough sets on the PIMA for the first time were applied by him and used the equal frequency binning criteria for intervals and then he created reduces by using Johnson reducer algorithm and classified using the batch classifier with the standard/tuned voting method (RSES). The rules were constructed for each of the data around 10 randomizations of the PIDD training sets from above [10]. The tests sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 70% to 86% with a mean of 74% and 95% CI of (71%, 76%).

A clustering algorithm that is used for predicting diabetes based on graph b-colouring technique was developed by Vijayalakshmi et al. Their experiments were compared to approach with K-NN classification and K-means clustering. The results showed that the clustering based on graph colouring is much better than other clustering approaches in terms of accuracy and purity. The proposed technique presented a real representation of clusters by dominant objects that assures the inter cluster disparity in a partitioning and used to evaluate the quality of clusters [11].

III. PROPOSED WORK

3.1 CKD analysis

Chronic Kidney Disease or chronic renal disease gradually evolutions and usually after few years the kidney loses its functionality. In general, it may not be spotted before it loses 25% of its functionality. In the initial stage of renal failure may not be predictable by the patients since kidney failure may not give any symptoms primarily. Kidney failure treatment targets to govern the causes and slow down the advance of the renal failure. If treatments are not adequate, patient will be in the end-stage of renal failure and the last treatment is dialysis or renal transplant. At present, 4 out of every 1000 person in the United Kingdom are suffering from renal failure [12] and more than 300,000 American patients in the end-stage of kidney disease survive with dialysis [13]. Due to detecting the chronic kidney failure is not feasible until the kidney failure is entirely progressed; thus, realizing the kidney failure in the first stage is enormously important.

Through early diagnosis, the act of each kidney can be taken under control, which leads to decreasing the risk of irreversible consequences. For this reason, routine check-up and early diagnosis are crucial to the patients, for they can prevent vital risks of renal failure and related diseases [12]. Therefore, it can be distinguished by measuring factors, and physicians can decide treatment processes, reducing the rate of evolution [14]. The purpose of medical diagnosis is to mine useful information from the immense medical datasets which are accumulated frequently [15]. Enormous mainstream of the studies on medical datasets are related to cancer diagnosis.

Sharma et al. used Naïve Bayes classification algorithm-based classifier to predict the Diabetes in Indian population. The authors reported 76.30% accuracy which is higher than the other classifiers (Decision Tree, K-Nearest Neighbors, Random Forest and Support Vector Machines) [16].

3.2 Feature selection and extraction

Feature selection is the key arena in knowledge discovery, pattern recognition and statistical science. The persistence of feature selection is to eliminate a subset from inputs which are not significant. Features do not depend on information about predictive classes. Reducing the dimensions of features and unrelated features can produce an inclusive model for classification. The main challenge of feature reduction is recognizing the best subset of features to achieve the best results of classifications [17]. Feature selection can simplify the data realization, decrease over fitting problem and the size of data storage; also, it can decrease the cost of train to obtain higher accuracy [18]. Feature selection methods can be characterized into three groups. Filter, wrapper and embedded methods. Fig. 3.1 shows the schema of filter and wrapper methods for feature selection and classification algorithm for classifying the selected subset methods used in this paper.

Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290

The filter method chooses the features whose levels are the highest among them, and then the selected subset can be prepared for any classification algorithm. After applying feature selection with filter method, numerous classification algorithms could be estimated (Fig. 3.1) [19]. An appropriate feature selection yields to performance improvement of classifier by reducing the computing time and by using optimized data in the dataset [20]. Furthermore, Filter method is a popular method in feature selection due to its fast performance and scalability [21, 22]. Wrapper method estimates scores of feature sets that rely on the predictable power by using a classifier algorithm as a black box [23]. Assessment of specific subset is attained by performance test and training on that precise dataset.

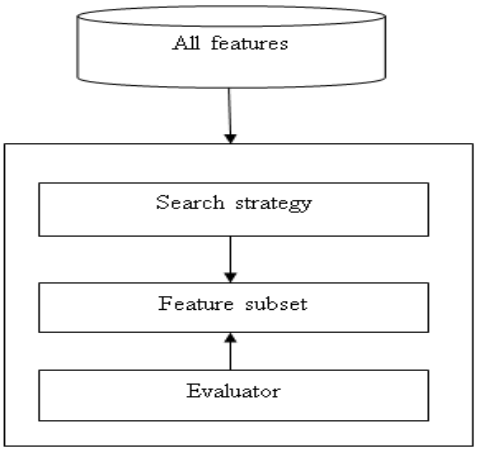

Figure 3.1 Schema of the filter feature selection method

The wrapped search algorithm around the classifier gains the space of all features of subsets [19]. 2n (n is the number of features), various assessments are required for a full search in the wrapper method. Although dealing with correlated features and finding the relevant associations are the advantages of this approach, it might lead to the over fitting problems [24]. The advantage of embedded method is that they interact with their classification model, and they do not have complicated computation. Decision support systems use different methods to reduce the features dimension and classification algorithms to diagnose several kinds of diseases.

3.3 WEKA classifier

Open-source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. If you detect the beginning of the movement of the image, you will recognize that there are many stages in dealing with Big Data to make it appropriate for machine learning.

- First, the raw data collected from the field. This data may contain some null values and unrelated fields. You use the data pre-processing tools provided in WEKA to cleanse the data.
- Then, you would save the pre-processed data in your local storage for applying ML algorithms.
- In ML model that you are trying to develop you would select one of the options such as Classify, Cluster, or Associate. The Attributes Selection allows the instinctive collection of features to create a concentrated dataset.

3.4 Decision tree and mining

Unwanted data are collected from the field contains things that leads to wrong analysis. The data may contain null fields; it may contain columns that are unrelated to the current examination, and so on. Thus, the data must be pre-processed to meet the requirements of the type of analysis you are seeking. This is done in the pre-processing module.

- Resource discovering and filtering: Data centre Broker discovers the resources present in the network system and collects status information related to them.
- Resource selection: The source is selected based on certain restrictions of task and resource.
- Task submission: Task is submitted to source selected. This is determining stage.

3.5 Result analysis

The datasets were pre-processed and classified using WEKA tool. Obtained the feasible results are found after classification and the proposed work maximizes the accuracy of finding optimal results.

IV. EXPERIMENTAL RESULTS

We have implemented our proposed strategy on the WEKA 3.8.4 simulator to appraise the practicability along the performance of supervised learning techniques (Fig. 4.1, 4.2, and 4.3). Disease prediction as well as management in large scale distributed system is complex. Therefore, for analysing the algorithm we use WEKA 3.8.4 simulators before applying them in real system.

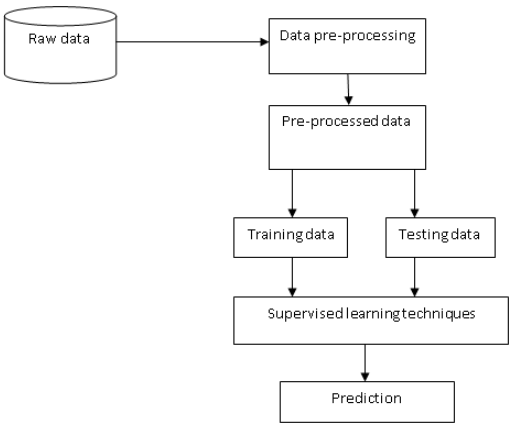

Figure 4.1 Architecture of prediction strategy

Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290

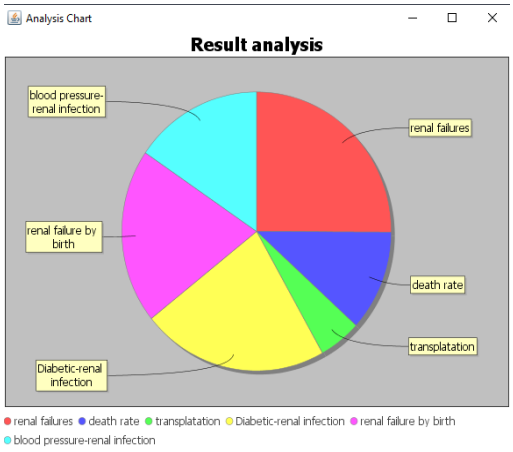

Figure 4.2 Analysis chart

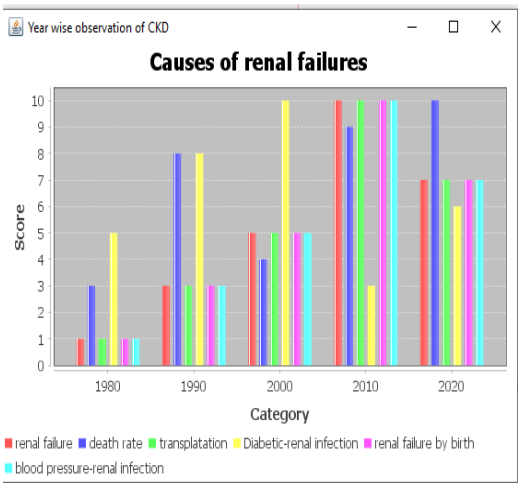

Figure 4.3 Year wise observation of CKD

V. CONCLUSION

5.1 Conclusion

Compared to related works, we have studied different supervised learning algorithms. We have analysed 14 different attributes related to CKD patients and predicted accuracy for different supervised learning algorithms like Decision tree and Identification of Pattern Mining. From the results analysis, it is observed that the decision tree algorithms give the accuracy of 91.75% and IPM gives accuracy of 96.75%. When considering the decision tree algorithm, it builds the tree based on the entire dataset by using all the features of the dataset.

The benefit of this organization is that, the prediction process is fewer overwhelming. It will help the doctors to start the treatments early for the chronic kidney disease patients and also it will help to spot more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a supervised learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.

5.2 Future research

This work will be considered as part of implementation for the healthcare system for CKD patients. Also, postponement to this work is that application of deep learning since deep learning provides high-quality performance than machine learning algorithm.

REFERENCES

[1] I.H. Witten, E. Frank, M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques* (United States: Elsevier, 2011).

[2] S. Varughese and G. Abraham, Chronic Kidney Disease in India, *Clinical Journal of the American Society of Nephrology*, 13(5), 2018, 802-804.

[3] H. Masethe and M. Masethe, Prediction of heart disease using classification algorithms, *Lecture Notes in Engineering and Computer Science*, vol. 2, 2014, 809–812.

[4] S. B. Patil and P. D. Scholar, Extraction of significant patterns from heart disease warehouses for heart attack, *International Journal on Computer Science and Network Security*, 9(2), 2009, 228–235.

[5] P. K. Anooj, Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules, *Journal of computer sciences*, 24(1), 2012, 27- 40.

[6] V. Manikandan and S. Latha, Predicting the analysis of heart disease symptoms using medicinal data mining methods, *International Journal of Advanced Computer theory and Engineering*, 3(2), 2013, 46-51.

[7] R. Chitra and V. Seenivasagam, Review of heart disease prediction system using data mining and hybrid intelligent techniques, *ICTACT Journal on Soft Computing*, 3(4), 2013, 605-609.

[8] A. G. Kkaregowda, V. Punya, M. A. Jayaram and A. S. Manjunath, Rule based classification for diabetic patients using cascaded k-means and decision tree c4.5., *International Journal of Computer Applications*. 45(12), 2012, 45–50.

[9] J. Han, J. C. Rodriguez and M. Beheshti, Diabetes data analysis and prediction model discovery using RapidMiner, *Second International Conference on Future Generation Communication and Networking*, 2008, 96-99. doi: 10.1109/FGCN.2008.226.

[10] J. L. Breault and Breault JL. Data mining diabetic databases: Are rough sets a useful addition. *In: In Proc 33rd Symposium on the Interface, Computing Science and Statistics*, 2001.

Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290

[11] D. Vijayalakshmi and K. Thilagavathi, An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis, *International Journal of Applied Mathematical Research*, 1(4), 2012, 520-530.

[12] T. K. Chen, D. H. Knicely and M. E. Grams, Chronic kidney disease diagnosis and management, *The Journal of the American Medical Association*, 322(13), 2019, 1294–1304.

[13] A. S. Go, G. M. Chertow, D. Fan, C. E. McCulloch and C. Y. Hsu, Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization, *The New England Journal of Medicine*, 351 (13), 2004, 1296-1305.

[14] R. Thomas, A. Kanso and J. R. Sedor, P. Kathuria and B. Wedro, Chronic kidney disease and its complications, *Prim Care*, 35(2), 2008, 329–vii.

[15] M. J. Huang, M. Y. Chen and S. C. Lee, Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis, *Expert Systems with Applications*, 32(3), 2007, 856–867.

[16] G. Sharma and U. Hengaju, Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes, *International Journal of Advanced Networking and Applications*, 12(1)2020, 4509-4518.

[17] R. K. Singh and M. Sivabalakrishnan, Feature selection of gene expression data for cancer classification: A review, *Procedia Computer Science, 50, 2015, 52–57.*

[18] S. Chao-Ton and C-H. Yang, Feature selection for the SVM: An application to hypertension diagnosis, *Expert Systems with Applications*, 34(1), 2008, 754–763.

[19] B. Kumari and T. Swarnkar, Filter versus wrapper feature subset selection in large dimensionality micro array: A review, *International Journal of Computer Science and Information Technologies, 2(3), 2011, 1048–1053.*

[20] O. Villacampa, *Feature selection and classification methods for decision making: a comparative analysis*, CEC Theses and Dissertations. College of Engineering and Computing. Nova Southeastern University, Florida, USA, 2015.

[21] A. G. Karegowda, M. A. Jayaram and A. S. Manjunath, Feature subset selection problem using wrapper approach in supervised learning, *International Journal of Computer Applications*, 1(7), 2010, 13–17.

[22] B. H. Cho, H. Yu, K. W. Kim, T. H. Kim, I. Y. Kim and S. I. Kim, Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods, *Artificial Intelligence in Medicine*, 42(1), 2008, 37–53.

[23] L. Ladha and T. Deepa, Feature selection methods and algorithms, *International Journal of Computer Science and Engineering*, 3(5), 2011, 1787–1797.

[24] L. Mousin, L. Jourdan, M. E. Marmion and C. Dhaenens, Feature selection using tabu search with learning memory: learning Tabu Search, *10th International Conference. LION 10. Ischia, Italy, 2016. doi:10.1007/978-3-319-50349-3_10.*

AUTHORS BIOGRAPHY

S. Rajarajeswari is a Research scholar in PG and Research Department of Computer Science, M.G.R College, Hosur, Tamilnadu. She has been involved in several international and national research conferences and presented various papers related to intrusion tolerance and network security, data mining. Her main research interests are: data mining, network security, and distributed systems.

T. Tamilarasi currently works as an Assistant Professor in PG and Research Department of Computer Science, M.G.R College, Hosur, Tamilnadu. Her area of interest includes, research in robotics, human–robot interaction, machine learning, deep learning Image processing and data-mining, and has published lot of research articles and chapters in book. She was a reviewer of Advances in Science, Technology and Engineering Systems Journal (ASTESJ).